# TOWARDS VISUAL FEATURE SELECTION FOR MULTIVARIATE TIME SERIES DATA

MASTER THESIS

submitted to the Faculty of Computer Science
at Otto-von-Guericke University, Magdeburg


LENA CIBULSKI

**Advisors**

Prof. Dr.-Ing. habil. Bernhard Preim
Department of Simulation and Graphics, Otto-von-Guericke University, Magdeburg

Dr.-Ing. Thorsten May
Fraunhofer Institute for Computer Graphics Research, Darmstadt

Magdeburg, November 1, 2017

**Lena Cibulski:**
*Towards Visual Feature Selection for Multivariate Time Series Data*
Master Thesis, Otto-von-Guericke University, Magdeburg, 2017.

# DECLARATION

I hereby declare that I have authored this master thesis independently and without the use of publications and resources other than those explicitly stated in the references.

*Magdeburg, November 1, 2017*

LENA CIBULSKI

## SCIENTIFIC ENVIRONMENT

The presented thesis has been submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science to the Department of Simulation and Graphics at Otto-von-Guericke University, Magdeburg.
I would like to thank *Bernhard Preim* for supervising this thesis, without whom I might not have started research in the field of visualization. From the beginning, he encouraged me to work scientifically and offered indispensable support along the way. I would also like to express my thanks for his valuable input and incredibly fast feedback that contributed to this thesis.

The thesis has been developed during a half-year stay at Fraunhofer Institute for Computer Graphics Research in Darmstadt under the guidance of *Thorsten May*. In this context, I would like to express my thanks for his on-site supervision and support. Besides offering constant motivation, he provided constructive ideas and new impulses regarding the working direction.

## ABSTRACT

Time series analysis and modeling are essential tools for the transfer of knowledge across time, also called forecasting. This often involves the task of identifying the least number of features that are most useful for building a model that accurately forecasts a target without suffering from dimensionality issues. This is challenging, because time series involve many different characteristics that need to be captured by a model. Traditional wrapper approaches are bound to the actual learning algorithm that builds the model, which requires computational effort and limits their range of application. Filter methods are independent of the future model, but mostly take the form of a black box algorithm, which does not allow analysts to monitor and interactively guide the feature selection. In this thesis, the filter concept for multivariate time series is advanced by making use of the human perception and interpretation abilities for independent evaluation of a feature subset's quality.

To ensure independence, we derive a quality criterion from a general assumption about the relationship between input and output in a valid model. An overview visualization enables analysts to visually assess its validity and to steer the analysis towards regions of interest, where the feature subset's quality is not sufficient. Critical regions can be analyzed in detail using the surrounding system of linked views. Findings contribute to an interactive refinement of the feature subset, which might also include the analyst's expertise. We evaluate the proposed method by applying it to real-world sensor data and an artificial time-oriented data set. The analyst was able to quickly distinguish well-explained regions from critical parts of the feature space, for which the identification of an additional explanatory feature could be tackled straight-away. Due to visualization constraints, the approach can handle only two-dimensional feature subsets, which are taken as input to perform one feature selection iteration. Still, it might be an inspiring step in the direction of universal dimension reduction that involves the human strengths.

# CONTENTS

# 1 INTRODUCTION

A time series is a set of data points indexed in temporal order. Multivariate time series arise in application areas ranging from the economic sector over meteorological and public health research to engineering. Time series analysis and modeling play a fundamental role for understanding and describing the relationships within a data set. A popular example of use is time series prediction, where future values of a target feature are forecast based on knowledge about the investigated system.

Knowledge manifests in relationships, which indicate explanatory power. A common statistical modeling approach that characterizes such relationships is regression analysis. We deal with data that is not only time-oriented, but also multivariate. Modeling therefore takes the form of multiple, univariate regression. Considering all available features might result in an unnecessarily complex model suffering from dimensionality issues like overfitting. For this reason, only truly relevant features, i.e. features that contain information about the target, should be included in the model, while redundancy is to be minimized. Feature selection aims at finding the minimal descriptive subset of features that together are most useful for explanation of the target. In this way, it contributes to more simplified and generalized models, cost-effective model building, and an increased model accuracy.

A time series might involve different characteristics like trends, seasonal patterns, or level shifts. All those characteristics need to be captured by a model, while keeping it as simple as possible. Selecting the most useful features to approximate the data generating process poses a challenge. Feature selection methods can be classified into two categories. Wrapper approaches require an execution of the actual model building algorithm, e.g. learning a classifier, to measure the quality of a feature subset. In contrast, filter methods are not tuned to the actual model generation. They filter out irrelevant features based on general relationships and are independent of the type and generating process of the future model. The approach developed in this thesis should not require the specification of a statistical model, which limits the feature selection to filter methods.

The aim of this thesis is to enhance the filter concept by combining it with visualization techniques that allow the analyst to monitor and steer the feature selection. To the best of our knowledge, relevant filter methods either include Interactive Visual Analysis techniques, but do not address time-oriented data, or they consider time series, but are implemented as automatic algorithms. We provide an interactive filter approach for numerical time-dependent data, which makes use of the analyst's visual perception for evaluation of a feature subset's quality. As the method does not rely on future model characteristics, it can be applied to regression problems, for which the model class is not known yet. This allows analysts to postpone the choice of the model class until the actual model fitting (not covered in this work). However, the model-free restriction limits the choice of quality measures.

We derive a quality measure from the general assumption that a valid model outputs the same predictions based on equal inputs. Given a feature subset, this assumption can be verified by comparing the target behaviors (predictions) that are associated with equal value combinations (inputs) in a set of training data. Interactive Visual Analysis enhances this comparison by enabling analysts to make use of their visual and interpretative abilities. The target behaviors are presented as a set of curves, whose variance provides information on the uncertainty of the feature subset. Domain knowledge can be efficiently included in the decision as to whether the feature subset might be adequate for prediction of the target. If it is not sufficient, the analyst can interactively steer the analysis towards critical regions. Information Visualization techniques then support an exploration of how the current uncertainty can be associated with other features, which might serve as a starting point for refinement.

This thesis addresses the following contributions regarding visual feature selection for multivariate, time-oriented data:

- Evaluate the quality of a given feature subset independent of the specification of an analytical model or model building algorithm.
- For this purpose, transfer concepts from event-based analysis of health records to numerical time series and regression modeling.
- Enable analysts to interact with the presented data to apply domain knowledge, identify critical regions, and refine the feature subset.
- Apply the method to one real-world and one artificial time-oriented data set to examine its benefits and limitations.

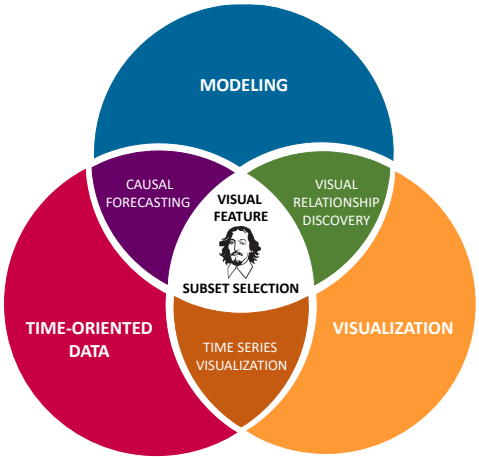



**Figure 1:** *The thesis' core topic addresses the intersection of three research areas.*

The topic of this thesis primarily intersects with three research areas (Figure 1): (1) the problem definition is taken from statistical modeling, (2) the time-oriented nature of the underlying data set requires appropriate methods, and (3) the presented approach relates to the field of Interactive Visual Analysis of high-dimensional data.

Chapter 2 and 3 give an overview of scientific work performed in the context of the named research fields. Chapter 2 covers fundamental concepts and challenges related to statistical modeling. Chapter 3 presents related work regarding Interactive Visual Analysis, focusing on time-oriented data, relationships in high-dimensional data, and feature selection.

Chapter 4 summarizes the contributions of this work.

Chapter 5 presents the visual feature selection approach. It is independent of the model class and generating process. Evaluation of a feature subset involves general assumptions about the modeling problem and concepts adapted from the health care domain. The main visualization is linked to other views that allow for an analysis from different perspectives.

Chapter 6 and Chapter 7 provide an informal evaluation of the proposed approach. They present a case study from the domain of vehicle dynamics as well as an exploratory analysis of an artificial data set.

Chapter 8 and Chapter 9 give a summary of the presented work and outline remaining challenges to be addressed in the future.

# 2 STATISTICAL BACKGROUND

## 2.1 FUNDAMENTALS OF PREDICTION

On a basic level, a *prediction* is nothing but a statement about an event that might happen in the future. Predictions come in different shapes that range from weather forecasts and the risk of developing a disease to the prognosis of a stock's future value or the outcome of a sporting event. Predictions are associated with uncertainty [17]. This is reflected in confidence intervals that are computed to assess the certainty in predicting a continuous variable or in probability values that are computed for outcomes of a discrete variable. Still, predictions can be useful for planning and decision-making in a wide range of applications, whether it be in physics [79], economics [62], politics [68] or the healthcare sector [27].

A predictive statement is often derived from *knowledge* about the underlying data. For example, in health care applications, conclusions about associations between elevated laboratory values and death rate in patient cohorts are drawn, based on which the mortality risk is predicted [27]. There are various ways, in which knowledge can be inferred, but most empirical methods lead back to predictive modeling. Methods for knowledge discovery evolve from diverse research areas like machine learning, neural networks, pattern recognition, or statistics [43].

In the context of this work, prediction can be viewed as a part of *inferential statistics*. Statistical inference uses patterns in observed data to draw conclusions about the more general population, from which the sample was drawn [10, 17]. Let us consider the relationship between father's height and son's height. We would expect taller fathers to have taller sons and shorter fathers to have shorter sons. Obviously, we cannot investigate the entire population to test this hypothesis. Inferential statistics allows for a statement about the unobserved population based on an observed sample of father–son pairs. The sample needs to accurately represent the population of interest. Not only the statement itself is associated with uncertainty, but also the underlying data. For some father–son pairs fatherhood might not be guaranteed, leading to outliers that can distort the inference.

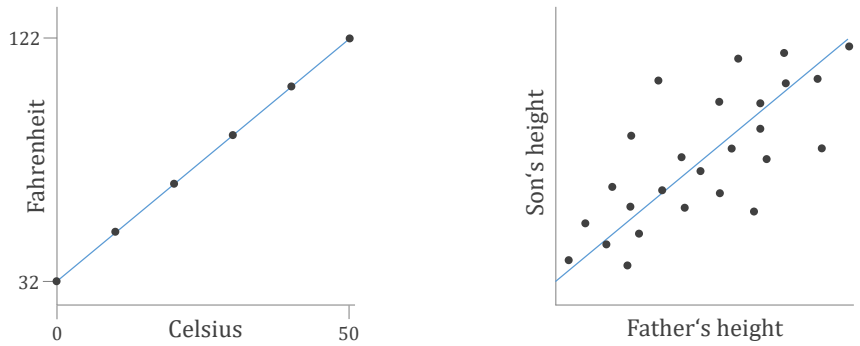



**Figure 2:** *Two types of relationships: a functional relationship is an exact relationship (left). A statistical relationship includes scatter and represents a tendency between two features (right).*

As the example suggests, we focus on inferences in the form of *statistical relationships* among involved features. A relationship is given between a *dependent* feature $Y$, i.e. the feature to be predicted, and one or more *independent* features $X_i$, on which the values of the dependent feature are hypothesized to depend. Unlike a functional relationship (Figure 2, left), e.g. the conversion between temperature in degrees Celsius and degrees Fahrenheit, a statistical relationship represents a discernible, but not necessarily exact relation between two features (Figure 2, right). If the characteristics of a relationship are known, unobserved values of the dependent feature can be predicted based on the values of the independent features.

When dealing with the field of statistical modeling, one will come across the terms relationship, dependence, association, and correlation. These terms also have their use in other research areas and even in everyday language. The difficulty with these terms is that (1) the same terms are used for different concepts and (2) different terms are used interchangeably. Their meaning highly depends on the usage context. For these reasons, we provide individual definitions, which apply within the context of this thesis. They are inspired by the terms' common usage throughout literature. The majority of descriptions in literature puts association on a level with dependence [52], which in turn is used interchangeably with relationship. Therefore, we only define the terms statistical relationship and correlation.

**Definition 0.1 (Statistical Relationship)** *A symmetric – but not necessarily causal – relation, where two or more features are statistically dependent, i.e. information about one feature are relevant for the assessment of another feature.*

A formal definition of the term independence between two features X and Y is: "X is independent of Y, if the distribution of X given $Y = y$ does not depend on the value y of Y" [47]. We define a statistical relationship as a condition, where the mathematical property of independence is not satisfied. Consequently, when two features are statistically related, one feature contains information about the other feature. The value of one feature is *in some way* connected to the value of the other feature.

**Definition 0.2 (Correlation)** *A measure that quantifies the strength and direction of a statistical relationship.*

This definition follows the common usage of correlation as a measure of the strength and direction of a relationship, which are not known for a statistical relationship. The majority of publications in statistics define correlation as a measure of linear dependence [20, 46, 50], rather than also considering non-linear relationships [84].

Investigating the behavior of a dependent feature Y based on changes of the independent features $X_i$ can provide useful information about how their relationship can be characterized. Table 1 illustrates all four possible combinations of changes in Y and $X_i$ with binary assignment yes or no. If Y depends on $X_i$, we would expect Y to change as the $X_i$ are varied (Q1). The same holds for the diagonally opposite combination, where Y does not change as a result of the $X_i$ being constant (Q4). If Y stays the same, although the $X_i$ change (Q2), this is unexpected behavior, as opposed to Q1. Again, assuming that Y depends on the $X_i$, we would expect that changes in $X_i$ affect Y and cause it to change. Consequently, if this is not the case (i.e. Y stays constant as for Q2), this indicates that the dependence assumption does not hold. Another kind of unexpected behavior arises, when the $X_i$ do not change, but Y changes (Q3). This case does not provide any indication concerning the assumption's correctness, but simply states that the $X_i$ not sufficiently explain the dependent feature Y.

| | | Y Changed | |
|---|---|---|---|
| | | Yes | No |
| $X_i$ Varied | Yes | Expected (Q1) | Unexpected (Q2) |
| | No | Unexpected (Q3) | Expected (Q4) |

**Table 1:** *Expected and unexpected behavior under the assumption that the dependent feature* Y *in some way depends on the predictors* $X_i$.

The characterization of correlation is based on the statistical model that describes the given data set. As mentioned in Chapter 1, we intend to postpone the choice of a specific statistical model for as long as possible. Therefore, in this work, we provide a model-free approach to the identification of relationships between features. For our objective of determining a set of features with a high predictive power, it is sufficient to identify and compare dependences. This only applies to the determination of features, not for the prediction step itself. For the feature selection, we do not need to quantify the strength of relationships or the predictive power, because the actual prediction accuracy does not play a role before optimizing the statistical model.

## 2.2 PREDICTION OF TIME-ORIENTED DATA

When it comes to data that is related to time, statistical methods that are not designed to consider temporality might reach their limits. Prediction in the sense of transferring knowledge about a data sample to the entire population is not necessarily the same as transferring knowledge across time. The latter, i.e. prediction in the context of time-oriented data, is called *forecasting* [17]. Most commonly this involves the analysis of trends.

Based on the model types involved, Makridakis and colleagues have divided quantitative forecasting methods into two categories: (1) *extrapolative forecasting*, and (2) *causal forecasting* [53]. Extrapolative forecasting methods predict future values as a direct function of historical data pattern. Taking the time series itself as a model, a feature's value is estimated beyond its observation range, assuming that the trend will continue without distortion. Causal forecasting methods take factors into account that are assumed to influence the feature to be forecast. Consequently, they are based on relationships between independent features and the target.

There are different scenarios, in which data related to time might be predicted. A scenario can be characterized by the number of independent features, the number of considered past time points, and the number of values to be predicted (Table 2). Based on a data table where the columns are features and the rows store time points, Figure 3 highlights the data table components that are involved in the different prediction scenarios.

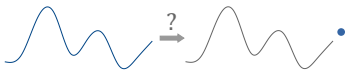
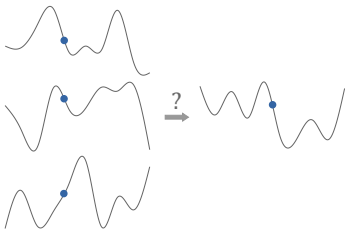
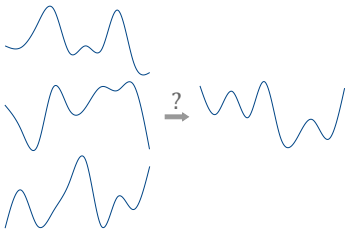
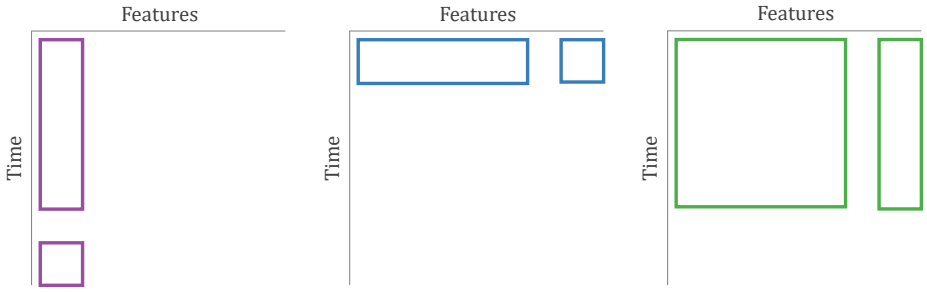
| | Features | Time points | Predictions | |
|---|---|---|---|---|
| Extrapolative Forecasting | one | many | one |  |
| Regression | many | one | one |  |
| Time Series Prediction | many | many | many |  |

**Table 2:** *Different scenarios involving the prediction of time-related values. Items colored in blue indicate input and output of the prediction. Grey, dashed lines represent the underlying temporal developments of independent (left) and dependent (right) features for reference.*




**(a)** *Extrapolative Forecasting* **(b)** *Regression* **(c)** *Time Series Prediction*

**Figure 3:** *Schematic representation of the data table components involved in the different scenarios shown in Table 2. Extrapolative forecasting involves one feature (a), regression is based on one time point (b), and time series prediction considers both multiple features and time points (c).*

#### SCENARIO 1 – EXTRAPOLATIVE FORECASTING

In most application areas, forecasting describes the process of predicting the value of a feature at some specified future date based on its past values. As an example, a stock's value might be predicted for the following day based on its development in the past 30 days. Therefore, one future value (i.e. at the following day) of one feature (i.e. the stock) is predicted based on its values at multiple past time points (i.e. the past 30 days). This is schematically displayed in Table 2, first row. As shown in Figure 3a, one column (i.e. feature) of the data table is considered along multiple rows (i.e. time points) to predict an additional row of the same column.

#### SCENARIO 2 – CAUSAL FORECASTING: REGRESSION

In epidemiology, a population is repeatedly observed to investigate the relation between an exposure to potential risk factors and the risk of actually developing a disease [81]. When taking only a cross section, i.e. a snapshot of the population at one point in time, one target value (e.g. the risk of developing a disease) is predicted based on multiple features (such as socio-demographic factors, genetic conditions, and medical status) at the same point in time. We already refer to this scenario as regression (which is introduced in Section 2.3), because it represents the same idea: exploring the relationship between a dependent feature and multiple independent features. The scenario is schematically depicted in Table 2, second row. Figure 3b shows that it corresponds to considering one data table row (i.e. one time point) over multiple columns (i.e. features) to predict another column in the same row.

#### SCENARIO 3 – CAUSAL FORECASTING: TIME SERIES PREDICTION

The problem statement as addressed in this work refers to the prediction of an entire time series, rather than a value at a specific point in time. Given are various independent features, e.g. different sensors, each of which measures its quantity over time. From this set of time series (i.e. one per sensor), the temporal development of a target feature is to be predicted. This case represents a mixed form of the two scenarios presented above: the target feature is predicted based on multiple past time points as realized in Scenario 1, but also based on multiple features as in Scenario 2. The schema in Table 2, third row depicts this scenario. In Figure 3c, we can see that this intuitively corresponds to extending the regression scenario (Figure 3b) along the data rows. However, simply extending this scenario would mean to consider the rows independently, i.e. the target values for each row are obtained by repeatedly applying Scenario 2.

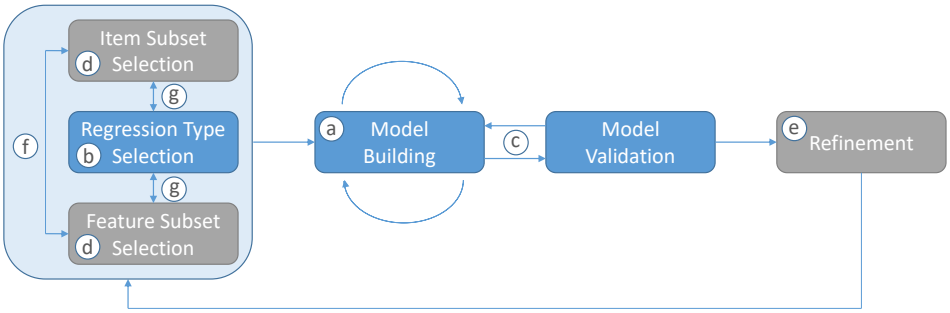



**Figure 4:** *The regression pipeline. Grey components are not mandatory.*

In contrast to that, this scenario takes the features' time-dependency into account: the value $Y(t)$ of the target feature $Y$ at time $t$ might not only be determined by the values $X_i(t)$ of the independent features at the same time $t$, but also by their values $X_i(t-1), X_i(t-2)$ etc. at previous time points. The most general form of a model can be described as follows:

$$Y(t) = F(X_{i_1}(t), ..., X_{i_k}(t), X'_{i_1}(t), ..., X'_{i_l}(t), Y'(t), t)$$

$X_{i_1}(t), ..., X_{i_k}(t)$ are the independent features whose values at time $t$ influence the target value, while $X_{i_1}(t), ..., X_{i_l}(t)$ are the features whose values at previous time points play a role. This is captured by considering their first-order derivatives, which becomes obvious when considering the backward difference quotient $X'_i(t) = \frac{X_i(t) - X_i(t-\Delta t)}{\Delta t}$.

## 2.3 REGRESSION ANALYSIS

*Regression analysis* is a commonly used statistical modeling technique for characterizing the relationships among quantitative features [15]. It helps to understand how a dependent feature $Y$ changes when the independent features $X_i$ are varied. In this sense, it can be used for forecasting based on knowledge that takes the form of correlation [10]. The main result of regression analysis is an explicit function of the independent features that models the relationship (Figure 4a). Note that the function class of the regression function has to be initially specified (Figure 4b), while it can still depend on unknown parameters. These regression parameters are then determined in a way that optimizes the fit of the regression function to the data (Figure 4c). The function can then be used to predict values of the dependent feature. This works best when a small number of features is considered together with large amounts of valid data [5].

Depending on the complexity of the given problem, the regression analysis process is not necessarily as straight-forward as described above. Additional steps, such as selecting an appropriate item or feature subset (Figure 4d) or a refinement according to the model validation results (Figure 4e), might be necessary to properly model the relationships contained in the data. The regression analysis process becomes even more complex, because these steps are highly inter-dependent. When a feature is added to the current feature subset, this might imply that the item subset needs to be adjusted as well (Figure 4f). On the other hand, if it becomes obvious that the regression type (e.g. linear) does not fit to the data set, the feature and item subsets might also need to be modified (Figure 4g).

### 2.3.1 *Causality*

In the context of this work, we are primarily interested in whether relationships can be identified at all, without considering their nature, i.e. whether they are causal or not. However, the concept of causality is an important aspect when dealing with statistical relationships. For completeness, we provide a brief overview of the potential and difficulties of integrating the concept of causality into regression analysis.

Cox and Wermuth suggest three broad notions of causality [18]. In this context, they state that a feature C is a cause of another feature R, when their relationship is not affected by considering additional sources of dependence in the form of other features that are explanatory to C. We refer to this definition of *causality* throughout this thesis. Depending on the feature composition, regression analysis is to different degrees helpful to assess a causal relationship [18]. The absence of causality for relations involving a common explanatory feature (Figure 5, left) can be identified. Such relationships are called symptomatic, because the seeming causality is induced by a common explanatory feature. In contrast, regression analysis does not allow for conclusions about the causality of relationships containing intermediate features (Figure 5, right).

The general problem with regression analysis failing to assess causality is the fact that *correlation does not imply causation*. When two features are correlated, it is all too intuitive to assume that one feature causes the other one. Many examples have been presented that raise awareness for such logical fallacy [31, 51, 80, 82]. Different works raise awareness for circumstances, under which experts tend to overestimate the predictability [5, 72]. Daniel Kahneman relates such overestimation to the underestimation

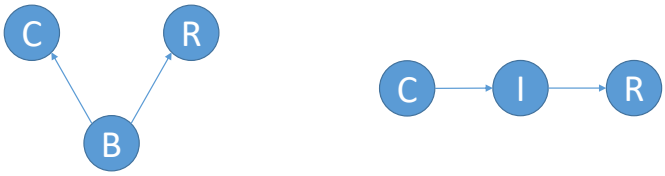



**Figure 5:** *Different notions of causality.* C *is not a cause of* R*, because the only dependence is that induced by their both depending on* B *(left). Using* I *as an explanatory feature of* R *might not detect the indirect path of* C *causing* I *and* I *being a cause of* R *(right).*

of randomness [39]. Logical fallacies do not mean that correlation is not at all related to causation. Tufte suggests the following statement: "Correlation is not causation but it sure is a hint." [75]. Therefore, correlation must be carefully interpreted within the given context [61].

Assessing the spectrum from correlation to causation for time-related data poses an extra challenge. If two features exhibit similar autocorrelation structures, a regression analysis might result in a strong relationship although no explanatory power is given [64]. Another complication originates from the fact that an effect does not necessarily follow the cause immediately. Two characteristics occurring closely in time are a strong hint for causality, but in general an event might be related to other events occurring after any time period. Not to be aware of this issue might lead to misunderstandings, if the role of time is not correctly understood [64].

### 2.3.2 *Curse of Dimensionality*

As many other approaches from various domains, regression analysis is affected by the *curse of dimensionality*. This term was introduced by Richard E. Bellman [8] and refers to different phenomena in the context of analyzing high-dimensional[1] data. Regardless of the application domain, the common problem is the rapidly increasing volume of the feature space as the number of features increases. This results in the density of data samples decreasing exponentially. With a very large number of features, each data point can be considered an outlier. For any method that requires statistics, this might lead to severe problems, as the amount of data needed to support a statistical evaluation drastically increases.

1 We refer to high-dimensional data when hundreds or even thousands of dimensions are available.

In particular, multiple regression suffers from this phenomenon. Considering many independent features favors *multi-collinearity*. This refers to two or more independent features being strongly correlated. As one independent feature can be linearly predicted from the other(s), collinear features are exchangeable within the model. As a consequence, the regression coefficients should not be interpreted as the contribution of individual features to the overall explanatory power of the model. Furthermore, a regression model involving multi-collinearity is not stable, i.e. small changes in the training data lead to large changes in the regression coefficients [9]. Including collinear features in regression analysis can also lead to relevant features being masked by less relevant features [56].

When a large number of dimensions is considered, the risk of *overfitting* the regression model emerges. When the number of features is too high, the regression model consists of too many terms for the number of observations [6]. At some point, the predictive power reduces as the dimensionality increases. This is known as the *Hughes phenomenon* [34]. The model then becomes tailored to fit the unique quirks of the training sample rather than reflecting the relationships within the population. Consequently, the model is not generalizable.

Obviously, the solution to these problems is to reduce the dimensionality of a given problem. This is called *dimension reduction*. In this context, two questions arise: (1) What is the optimal number of features to be included in the regression analysis? and (2) Which features should be included? One class of approaches to dimension reduction are *feature selection* methods, which aim at selecting an optimal descriptive subset of the original features rather than generating synthetic features.

## 2.4 FEATURE SELECTION

This is the primary task associated with the objective of statistical modeling for time-oriented data. When performing a regression analysis in high-dimensional data space, not all features are actually important for the result. Feature selection uses the presence of redundant or irrelevant features to reduce the number of features in the data set without losing too much information. It is performed prior to modeling in order to avoid the consequences of the curse of dimensionality as described above, aiming at a simplification of regression models.

Focusing on the most promising features for prediction allows for a better generalization of the model, because less model parameters have to be estimated from the given data points [11]. In the end, feature selection means to find the least number of those features that are most useful for prediction, i.e. that most contribute to prediction accuracy. Note that relevance (e.g. according to some correlation coefficient) does not necessarily imply usefulness [26]. As an example, selecting the most relevant features might not be useful when many of them are redundant.

Different decisions have to be made during feature selection. Let us again take a look at Table 1. The combinations classified as unexpected behavior (i.e. Q2 and Q3) indicate that the chosen independent features $X_i$ are not sufficient to entirely explain $Y$. This can arise in two different forms. Either the $X_i$ are not at all suitable for prediction of $Y$, because there is no relationship. In this case, one or multiple $X_i$ should be removed from the feature subset. On the other hand, the $X_i$ might contain relevant information about $Y$, but there is some variance in $Y$ that cannot be solely explained by the current $X_i$. Then, one or more additional features should be considered to predict $Y$.

John and colleagues proposed two classes to characterize existing feature selection approaches, namely the *filter* and *wrapper* methods [38]. The classes differ in the way in which the quality of a feature subset is measured. Filter methods are based on general measures, such as the relationship with the dependent feature to be predicted. In contrast to that, wrapper and embedded methods contain a feedback loop between feature selection and actual induction. Induction means predictive modeling in this context, e.g. by learning a decision tree or neural network.

We aim at developing a feature selection approach that does not at all depend on the function family used for the following model building step (not covered in this work). Filter methods evaluate features without considering the impact that the resulting subset has on the performance of the induction algorithm. For this reason, they are the feature selection class of our choice, although wrapper methods are in general recommended over filter methods [44].

### 2.4.1 *Filter Methods*

With filter methods, the feature selection can be viewed as a preprocessing step prior to the model building step. It explores general patterns in the given data set to filter out irrelevant features. Filter methods are based on quality measures that characterize the relationship between the features and the target. Features that are the least interesting are eliminated. The remaining features form the feature subset for induction. Taking a look at different approaches, the challenges with filter methods become evident.

The FOCUS algorithm proposed by Almuallim and Dietterich [3] selects the minimal subset of features that perfectly discriminates among the available classes. This might lead to generalization problems, i.e. when an ID feature is contained in the selected subset. Kira and Rendell presented the RELIEF algorithm [42], which evaluates a feature based on its correlation to the target feature. This evaluation does not consider the relationships between the selected features and thus tends to result in a subset containing redundant features. Quality measures have also been advanced to consider such redundancy [12]. Cardie used a decision tree approach for feature selection [14], where features that did not appear as splitting attributes in the tree are removed from the subset. However, features that are relevant for decision tree induction are not necessarily useful if another induction algorithm is chosen.

The separation of feature selection and induction also offers advantages. It is effective in terms of computation time, because the evaluation of a candidate subset does not require an execution of the induction. It is also relatively robust with regard to overfitting. For our work, the major advantage of separating feature selection and actual induction is the independence towards the induction algorithm. This allows filter methods to be combined with any kind of induction algorithm, whether it be a naïve Bayes classifiers or a Bayesian network. Even more important, it enables us to develop an approach that is totally independent of the chosen regression model. Thus, the choice of a model class can be postponed.

# 3 STATE OF THE ART

In Chapter 2, we have introduced fundamental principles of statistics as a background for the problem definition. The following chapter addresses the visualization context necessary for the method that is proposed in this thesis. We first give a short introduction into the concept of Interactive Visual Analysis, followed by issues to be considered when visualizing time-oriented data. To conclude this chapter, we put interactive visualizations into the context of relationship discovery and feature selection.

## 3.1 CONCEPT OF INTERACTIVE VISUAL ANALYSIS

In view of rapidly increasing amounts of data that are automatically recorded via sensors and monitoring systems, extracting valuable information is difficult [40]. Not being able to adequately explore the sheer amounts being generated, the data becomes useless for hypothesis generation and decision-making. Automatic filtering and data analyses yield reliable results when used for well-defined problems. However, for hypothesis generation or in-depth interpretation of analysis results, the ability to understand the procedure from data to results is crucial [41]. When dealing with complex data, combining the strengths of human and automatic data processing allows for analyzing data in a most effective way [40]. This is realized in *Interactive Visual Analysis* (IVA), a concept established by Helwig Hauser. It involves a step-wise analysis procedure: from descriptive statistics for each feature, over the presentation of correlations between multiple features to more advanced investigations if required.

The challenge with large and often high-dimensional data is to present them in a visual form that supports analysts in gaining insights and reduces the cognitive load. *Information Visualization* (IV) techniques are commonly used for the identification of patterns, trends, or correlations among features in multivariate data [19]. Interactive exploration benefits from simultaneously investigating different visual representations. Such methodology is supported by *Coordinated Multiple Views* (CMV), which are employed for various types of information [22, 54, 60]. For more detailed information on CMV, see the survey provided by Roberts [67].

CMV are commonly used to implement an *overview and detail* strategy, following Shneiderman's information-seeking mantra "Overview first, zoom and filter, then details on-demand" [71]. This strategy enables analysts to concretely analyze patterns in detail, which were observed in an overview. When viewing overview and detail representations side by side, analysts can consider multiple scenarios, compare different perspectives, and go back and forth between different exploration paths [67].

As Keim and colleagues state in their work, the user should be the supervising authority steering the analysis procedure according to his task [41]. Interaction techniques enable analysts to adjust the visual representations to his needs by filtering or selecting elements. Linking between multiple visualizations is performed via *brushing*, where all views update their content according to the selection of a data subset. Different brushing techniques are summarized by Roberts [67].

Brushing as an interaction technique is closely related to the *Focus and Context* (F+C) approach that has been developed in the IV domain, where it originally made use of different enlargement factors for different parts of the data. In combination, both brushing and F+C are used to draw the analyst's attention to a subset of data that is promising and thus should be analyzed in more detail. F+C techniques allow the analyst to investigate details on a subset of data, while still viewing a general overview as context for orientation within the same view. To distinguish between data in focus and context information, Hauser describes different visual characteristics, such as space, opacity, color, and frequency [28].

## 3.2 VISUALIZING MULTIVARIATE, TIME-ORIENTED DATA

Time, together with the three spatial dimensions, characterizes the four-dimensional space of the world we live in [2]. Every measurement is in one way or the other related to time and might only be meaningful in its temporal context. A deep understanding of temporal relations allows to learn from past developments and predict the future. There is no formal definition, but the majority of attempts to define *time orientation* include the idea that temporal aspects of the data are of central importance. Time-oriented data arises in various domains: from industry over meteorology and the finance sector to medicine. Typical tasks range from analyzing trends and repetitive patterns over identifying correlations to predicting future behavior [83]. Accordingly, a large number of visualization methods for time-oriented data in different applications have been published.

### 3.2.1 *Categorization of Time Representations*

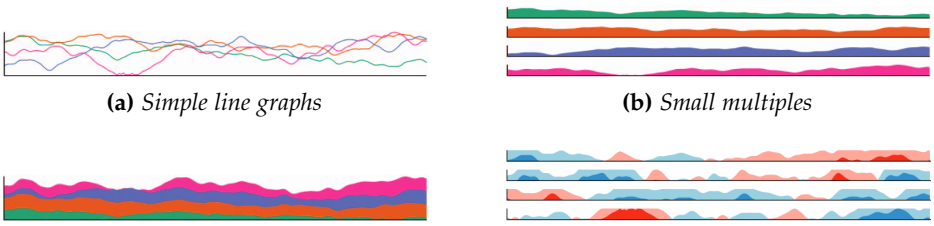
To address the diversity of proposed visualizations for time-oriented data, Aigner and colleagues have developed a categorization scheme covering the key characteristics of visualization methods [1]. Following this categorization scheme, we will introduce the fundamental settings of our approach by classifying it. Basically, the authors propose three categorization criteria (for a detailed description see [1]):

1. Time Axis
    - Primitives: **time points** vs. time intervals
    - Structure: **linear** vs. cyclic
2. Data
    - Frame of reference: **non-spatial** vs. spatial
    - Data type: **quantitative** vs. qualitative
    - Number of features: univariate vs. **multivariate**
3. Representation
    - Time dependency: **static** vs. dynamic
    - Dimensionality: **2D** vs. 3D

Now that we have introduced the idea of time and its role for data analysis, we will describe different approaches for visualization of time-oriented data. Because time-oriented data involves so many different aspects, we will focus on a number of selected techniques, whose categorizations are similar to that of our approach.

### 3.2.2 *Time Series Plot and Extended Variants*

As Tufte states, "the time-series plot is the most frequently used form of graphic design" [74]. A *time series plot* displays the change of a variable's values, while keeping the temporal ordering. In other words: observations of a variable are plotted against time. The oldest known example to representing changing values graphically dates from the tenth or eleventh century [21]. Once the numbers recorded at different time points are put into a graphical representation, trends and patterns can be identified. A time series plot is intuitive to interpret, because it addresses the natural perception of time as a linearly proceeding dimension.



**(a)** *Simple line graphs* **(b)** *Small multiples*

**(c)** *Stacked graphs* **(d)** *Horizon graphs*

**Figure 6:** *Different visualization techniques showing the same time series [37].*

The time series plot has been extended to multivariate data, yielding a number of variants addressing different tasks. Javed and colleagues give an overview of different ways to display multiple time series [37]. Plotting multiple time series with a common baseline is the most straight-forward way to compare different variables (Figure 6a). To reduce visual clutter, labels and axis ticks are omitted, resulting in what Tufte calls *sparklines* [76]. The time series can also be displayed one below the other, which additionally allows to color the area below the line to simplify the identification of individual graphs (Figure 6b). To save vertical space, the small multiples can also be stacked, such that each time series uses the value of its predecessor as a baseline (Figure 6c). However, the limited visual clutter of such *stacked graphs* comes at the expense of a more demanding comparison across time series. Another visualization, which keeps the visual clutter and space allocation low, are *horizon graphs* [66]. The approaches presented by Javed et al. [37] can be of use for perceiving simultaneous occurrences of certain features, e.g. peaks, across a number of variables. Such occurrences might indicate a relationship, where changes in one variable trigger changes in another variable. Instead of arranging time series plots in a vertical layout, they can also be set out in a radial layout, resulting in a visualization called *MultiComb* [73].

### 3.2.3 *Exploring Patterns of Evolution*

Time series plots as such depict the values of features at different time points. The change in individual features needs to be read from the gradient of the corresponding curve, which requires cognitive effort. Havre and colleagues present *ThemeRiver* [30] to visualize changes of theme occurrences in speeches, interviews, or articles over time. One theme is represented by a colored ribbon flowing horizontally with time. For each time point, the ribbon's width is mapped to the relative strength of the

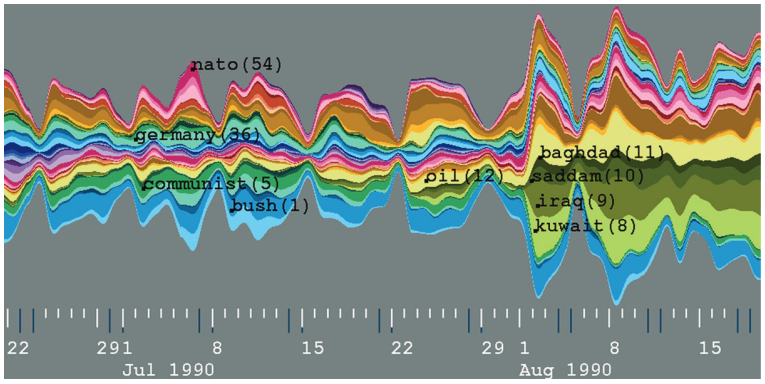



**Figure 7:** *ThemeRiver: a technique to visualize theme changes over time [30].*

theme at that time. Thus, changes in the strength of themes are perceivable with little effort by observing the varying width of the ribbons. To display multivariate data, the ribbons are stacked (Figure 7). ThemeRiver is highly useful for an investigation of general trends. It could also be used to depict simultaneous changes in the independent features and the target. A simultaneous widening and narrowing of multiple ribbons might indicate a relationship between the corresponding features. Interactive re-ordering of the ribbons enables the analyst to compare features without distortions, e.g. for investigation of a hypothesized relationship.

Another approach that is closely related to our method is proposed by Bach et al., who call it *Time Curves* [7]. Time Curves result from folding a time-line visualization into itself, such that similar time points are placed close together, while preserving the temporal order (Figure 8). Originally, Time Curves were used for exploring patterns of evolution, like slow progression, sudden changes, or reversal to a previous state, in temporal data originating from document histories. The authors claim that this metaphor can be applied to any data set where a similarity metric between time points can be specified. We use a special case of this metaphor

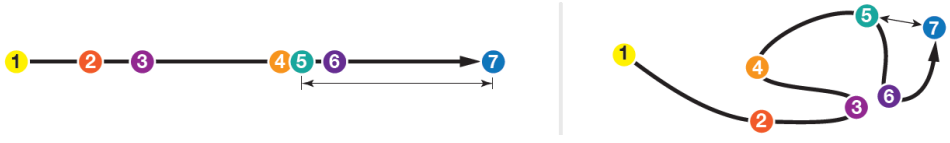



**Figure 8:** *Folding time: the time-line (left), where similar colors indicate similar time points, and the resulting time curve (right), where spatial proximity indicates similarity. Image adapted from [7].*

for our synchronization approach. We define the similarity between two time points $t_1$ and $t_2$ as the Euclidean distance of the corresponding data points in a feature space $X_1 \times X_2$ of independent features. Folding the time-line then equals plotting the data points in a scatter plot with $X_1$ and $X_2$ on the axes and connecting them according to their temporal order. For a detailed description of its application see Section 5.4.3.

In the following, we will turn towards one of the main objectives of this work: the identification of relationships for the purpose of selecting a descriptive feature set. We present several approaches to a visual assessment of relationships, most of them are targeted at time-oriented data. Nevertheless, we also include methods that do not deal with time-oriented data, but are related to our work from a conceptual point of view.

## 3.3 VISUALLY SUPPORTING RELATIONSHIP DISCOVERY

As for many other analysis tasks, visualization can be used to identify relationships within data sets, whether it be for choosing the right parameters or predicting the outcome of a variable. The most basic multivariate visualization techniques that address the identification of multidimensional relationships are scatter plots and scatter plot matrices [16] as well as parallel coordinates [36]. Amar and Stasko discuss the role of Information Visualization for higher-level tasks beyond the representation of data, such as decision-making [4]. They argue that IV techniques often lack the potential to overcome the gap between the perception of a relationship and building up confidence in the relationship by assessing its usefulness. They claim that systems should be built to support decision-making rather than leaving it to the experience of users.

### 3.3.1 *Relationship Exploration in Time-Independent Data*

Various approaches have been proposed for the discovery of patterns in large, multivariate data sets, where the temporal context does not play a role. Most of them were developed to support the identification of meaningful features for further analysis tasks, such as model building.

#### EXPLORING RELATIONSHIPS FOR DIMENSION REDUCTION

Krause and colleagues developed *SeekAView*, a Visual Analytics system that guides analysts in interactively building and refining subspaces out of high-dimensional data [49]. Analysts can either start the analysis by investigating system-generated subspace suggestions or by interactively

exploring relationships between dimensions themselves. The system can be asked for support on further steps whenever needed. SeekAView focuses on linear relationships, which contradicts our intention of building a system that does not rely on a selected class of regression functions.

A similar framework for dimensionality reduction and analysis, *DimStiller*, is provided by Ingram and colleagues [35]. It combines variance or correlation measures with visualizations like scatter plots to emphasize the underlying dimensions and their relationships. In this sense, it aims at answering questions concerning the meaningfulness of dimensions, the relationships between dimensions, and the validation of clustering with a given input dimensionality.

Although both systems cover the discovery of correlation patterns, they were primarily built for dimension reduction in all its many aspects. They focus on reducing the dimensionality of a data set using automated algorithms like Multidimensional Scaling, Principal Components Analysis, or Subspace Clustering rather than visually assessing a subspace's explanatory power for a target dimension. Subspace suggestions are not limited to original dimensions, but also include newly generated ones. In this sense, both frameworks focus on exploring relationships for the purpose of feature extraction rather than feature selection. In contrast to that, we only select dimensions from the features available in the data set.

#### FEATURE AND ITEM SUBSET SELECTION

In this sense, the *Rank-by-Feature Framework* provided by Seo and Shneiderman [70] is more closely related to our work, because it solely focuses on subsets of the original dimensions. The framework provides guidance for a systematic visual exploration of dimensions and their relationships for the purpose of discovering important dimensions for further analysis. The analysis work flow includes both information visualization techniques as well as statistical measures. The latter can be used to rank (pairs of) dimensions in order to rapidly identify the most important dimensions or the strongest relationships according to the specified measure. Independent of the choice of interestingness measures, the ranking concept can be beneficial for feature selection, as it supports users in determining which of the remaining features to add next to the current feature subset. The framework is limited to the discovery of global characteristics, i.e. all visualizations and statistical measures refer to the entire data set.

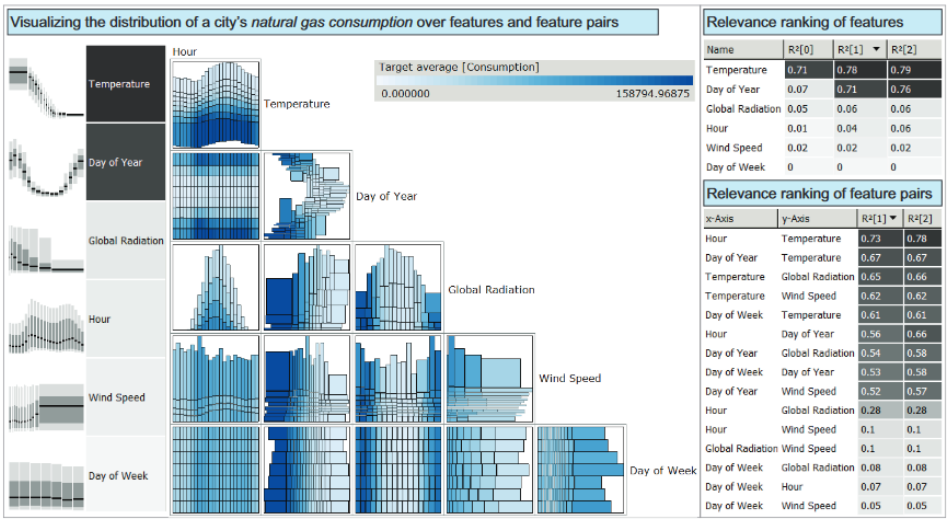



**Figure 9:** *The Partition-Based Framework for regression model building [57].*

Piringer and colleagues extend the framework to local characteristics, where univariate and bivariate statistical moments for ranking are computed with respect to a subset of the data. Data subsets can result from clustering or outlier removal. A more flexible way of defining them are brushes, which also enable a comparison of one subset to other subsets or the whole data set [65]. All views are linked, such that all visualizations and computations update, whenever a subset changes.

Regarding the objective of identifying meaningful dimensions as a starting point for tasks like model building, both approaches are closely related to our work. However, they are not designed to support regression-related tasks, which build upon relationships focusing on a target dimension. This makes it difficult to directly apply these methods to our task.

#### FEATURE AND ITEM SUBSETS FOR REGRESSION MODEL BUILDING

Guo et al. propose a visualization system for identifying linear trends between a dependent feature and independent features with the objective of interactively building explanation models [25]. Furthermore, the system supports users in navigating the model space and extracting data subsets fitting a given trend. It was also used to assess the uncertainty associated with a model. In contrast, our problem definition is taken from a stage prior to the model building step of the regression pipeline, where the model class and potential predictors are not known yet.

Mühlbacher and Piringer developed the *Partition-Based Framework*, which addresses different tasks related to building regression models [57]. In large parts, this involves the visual exploration of relationships between independent features and a quantitative target feature. For this purpose, the authors provide an overview of individual (1D) features and relationships for pairs of features (2D) (Figure 9). A ranking of relationships is employed for quantification. The Partition-Based Framework is closely related to our work, in that it addresses the exploration of relationships for use in regression analysis. By partitioning the feature space into data subsets, they even address the influence of item subset selection on the relationships between features. We also partition the feature space to achieve an overview visualization at an intermediate detail level, but do not consider particular item subsets for our analysis work flow. The key challenge that distinguishes our objective from that of the Partition-Based Framework is the time-dependency associated with our data, which is not addressed by Mühlbacher and Piringer. Another difference lies in their consideration of item subsets, which we do not provide.

### 3.3.2 *Relationship Exploration in Time-Oriented Data*

Adding time-dependency as a data characteristic to the task of exploring relationships in multivariate data for the purpose of forecasting poses an extra challenge. Approaches for time-oriented data range from general identification of relationships over simple extrapolative forecasting to effective sense-making using multiple features.

#### RELATING A TARGET TIME SERIES TO OTHER TIME SERIES

Zhao and colleagues provide a technique called *ChronoLenses* that allows for on-the-fly transformation of time series sections within a region of interest [83]. The relationship between time series can be investigated using the cross-correlation lens (Figure 10). However, a correlation within its focus area does not mean that the overall relationship is stable. By dragging the lens and therefore changing its focus area, the analyst can assess whether the correlation varies or exhibits a stable trend. We encounter a similar problem, because relationships can only be investigated for a part of the feature space. If a relationship was identified, it does not necessarily hold for the remaining feature space. Instead of employing interaction like dragging a lens, we arrange individual components to an overview visualization that depicts a relationship across the entire feature space.

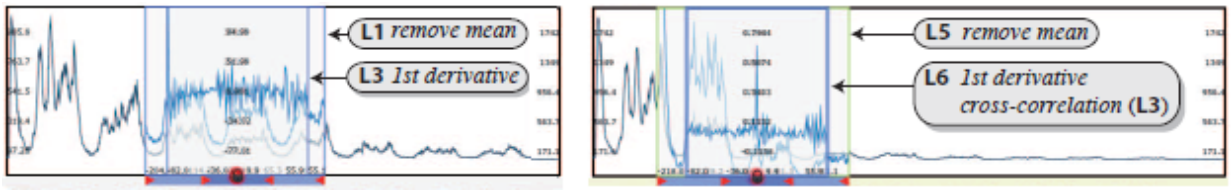



**Figure 10:** *ChronoLenses: analyzing the correlation between time series [83]. Lenses can be dragged along time to assess the correlation variation.*

Hochheiser and Shneiderman provide an interface called *Similarity-Based Forecasting*, which enables data-driven forecasting based on similar historical time series of the same feature [13]. For example, it can be used to forecast the closing price of an auction based on the price curves of past auctions selling the same object. For this purpose, they introduce the *River Plot*, an aggregation visualization depicting summary statistics of the past time series values. Its median is the forecast, while the remaining summary statistics indicate the forecast uncertainty. Similar to the River Plot, our approach also includes a visualization of time series associated with the target feature, whose variance characterizes the uncertainty of the forecast, i.e. the explanatory power of the considered feature subset. However, we visually assess the variance rather than quantifying it by a statistical measure.

Konyha and colleagues present a system for the interactive visual analysis of relationships in function graph ensembles [45]. A function graph $f(\mathbf{x}, t)$ is a time series that depends on a set of scalar independent features $\mathbf{x}$. It arises as output of a simulation run, where $\mathbf{x}$ is a set of control parameter values that it receives as input. An ensemble of function graphs then results from multiple simulation runs with varying control parameters. The objective that they address is two-fold: (1) determine how the shape of a function graph is related to combinations of scalar independent features (Figure 11, left) and (2) identify correlations between sets of function graphs (Figure 11, right), where a set results from multiple runs simulating the same feature. Both of these goals seem to correspond to our problem definition. However, in our case, (1) the independent features are function graphs themselves, which poses an additional challenge, and (2) we investigate the relationship between a target feature and multiple time-dependent features as two sets of function graphs, while their sets are equally sized and represent one feature each. Consequently, their techniques need strong adaptation to be applicable to our data set.

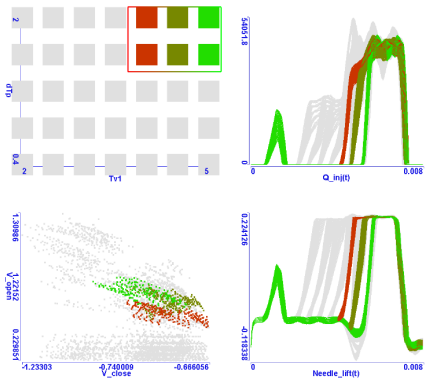


**Figure 11:** *Analysis of relations between control parameters (left) and time series (right) [45]. The sets of time series are correlated.*

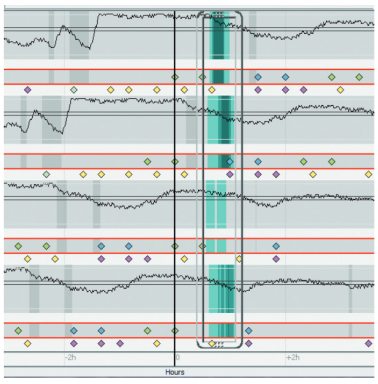


**Figure 12:** *CareCruiser: effects of a clinical action compared by vertical alignment [23]. A delayed drop of the parameter is revealed.*

### CAUSE-AND-EFFECT RELATIONSHIPS

The identification of relationships between time series might involve the search for items with a certain behavior, e.g. a strong increase, or event that anticipate changes in other time series within the data set. From such a finding, one might hypothesize that one feature triggers changes in another feature, thus indicating a relationship.

Gschwandtner and colleagues developed *CareCruiser* [23], a visualization prototype that supports analysts in exploring the effects of clinical treatment plans on a patient's condition. One of their objectives is the analysis of changes in a parameter curve, e.g. oxygen saturation, that might be caused by a clinical action, e.g. oxygen supply (Figure 12). This refers to cause-and-effect relationships between clinical actions and the temporal development of a patient's parameters. Immediate effects of applied clinical actions can be identified, which are a strong hint for causality due to the small time delay. The system also supports an alignment of actions below each other to enable a comparison of multiple patients based on the same clinical action. This is closely related to our synchronization approach, where we align time points exhibiting similar states (corresponding to a clinical action) to compare the behavior of a target feature (corresponding to the parameter curves). Our method differs from the described approach in that we do not deal with event-based, but continuous data. We need to assess whether a change in one time series might be related to or even causing any change in another time series.

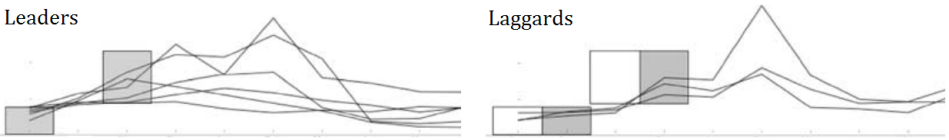



**Figure 13:** *Identifying items with similar trends at different times. Leaders are specified by a query (top). Laggards are identified by shifting the rectangles (bottom). Image adapted from [33].*

Hochheiser and Shneiderman address cause-and-effect relationships by a concept called *leaders and laggards* [33]. The leaders exhibiting the trend of interest are specified using selection rectangles (Figure 13, left). Laggards are then identified by the same number of selection rectangles, which are offset by one time period from their original counterpart (Figure 13, right). They only support the detection of laggards undergoing similar transitions to those of the leader, but at a later time. This does not cover the general hypothesis of a leader causing any change in its laggards.

## 3.4 FEATURE SELECTION USING INTERACTIVE VISUAL ANALYSIS

Feature selection is about finding a minimal subset of features, which is most useful for building a predictor [26]. Numbers primarily quantify the relevance of a feature or feature subset. However, relevance does not necessarily equal usefulness. Gathering together features with good individual predictive power does not mean that their combined predictive power is as high. As the search space of all feature subset candidates might be huge, Guo states that automated analytical processes are needed [24]. At the same time, such processes need to be tightly coupled with the human expertise in interpreting and evaluating patterns to guide the procedure.

### 3.4.1 *Interacting With Feature Selection Mechanisms*

Guo proposes an interactive feature selection method [24] that is based on a visualization of the relationships between 2D feature sets. A correlation matrix depicts a correlation measure for all 2D subspaces of the original feature space. For a better perception of interesting subspaces, the features are sorted such that correlated features are placed close to each other in the ordering. Subspaces that likely contain significant patterns are then visible as blocks of low correlation values and can be interactively selected from the matrix view.

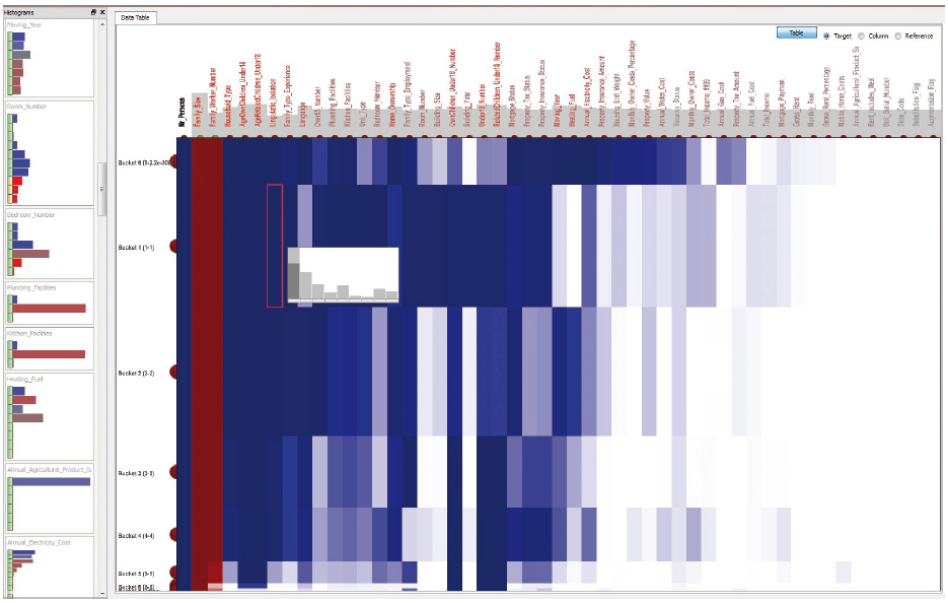


**Figure 14:** *SmartStripes interface for monitoring and steering feature selection [56]. The heat map depicts dependences between a selected (leftmost column) and remaining features. Dependence measures are partitioned w.r.t. entity subsets (rows).*

The coupling of data mining and visualization can also be realized by allowing direct interaction with a given feature selection algorithm through visualizations. May and colleagues present *SmartStripes*, a visual analytics system that enables the user to step into the feature selection process and retrace the interdependencies between feature and entity subsets [55, 56]. In this way, SmartStripes supports the monitoring of automatic methods and enables an interactive construction of feature subsets with a focus on entity subsets. In particular, the system was designed to be used with filter algorithms, which relates to the work presented in this thesis. In each iteration of the feature selection algorithm, the user can observe the quality measures computed for each feature. A heat map display shows the relation between a selected target feature and the remaining features in the data set (Figure 14). The rows of the heat map represent entity subsets. By investigating the subsets' individual contributions to the features' quality measures, the user can explore the interdependencies between feature and item subsets. If desired, the user can then directly steer the algorithm by choosing the next feature to be added according to the quality measures computed during the iteration. Item subsets that might skew the results of automatic feature selection techniques can be excluded from

the analysis. In this way, domain knowledge can be incorporated into the selection process. Scalability is addressed by sorting the features according to their quality as a candidate and displaying the most important features on the screen. The remaining features can be viewed by scrolling.

The SmartStripes approach is closely related to our work. Their similarity manifests in both approaches being independent of the actual feature selection algorithm and not being influenced by specific types of data or relations. Both also do not take the dependence of arbitrary features into account, but focus on relations between one dependent feature and the remaining independent features. In contrast to SmartStripes, we do not focus on the interdependencies between feature and entity subset selection, but instead aim at incorporating our features' time-dependence into the feature selection method.

Inspired by SmartStripes, Mühlbacher and Piringer allow for an investigation of local patterns when exploring the relationships between a target feature and a number of independent features [57]. Local relationship structures are made visible by partitioning the domain of involved features, which is also integrated into their interactive feature subset selection mechanism. The work flow consists of iteratively adding features, while ensuring that the resulting feature subset is valid in terms of redundant features and domain knowledge. Each iteration aims at reducing the remaining variance of the prediction, which is given by the deviation of the predicted values from the values of the training data. Generally, the described work flow matches the feature selection procedure considered in this work. However, it differs from our intentions in that it requires an evaluable regression model to determine the remaining variance. We also use our available data as training data to evaluate the current feature subset, but we characterize the remaining variance without the need of a model being given. The user solely visually determines whether the remaining variance is acceptable and, if not, selects an additional feature that is assumed to reduce this variance. Instead of building an initial model based on automatically determined features, we initialize our feature selection using a pair of features that is assumed to be useful based on domain knowledge. To conclude, this approach is highly similar to our work, except for the fact that feature subsets are evaluated using concrete regression models, which we do not want to presume.

| Method | Model-free | Item subsets | Domain knowledge | Feature ranking | Feature transformations | Time dependence |
|---|---|---|---|---|---|---|
| Interactive Feature Selection [24] | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| INFUSE [48] | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ |
| SmartStripes [56] | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| Partition-Based Framework [57] | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ |
| Our approach | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ |

**Table 3:** *An overview of reviewed approaches to interactive feature selection. None of them addresses time-oriented data (rightmost column).*

# 4 CONTRIBUTIONS OF THE THESIS

In this chapter, we briefly summarize the contributions of this thesis. Table 3 presents an overview of the selected approaches reviewed in the previous section. General functionalities like domain knowledge are supported by most of them, while aspects like feature transformations or item subsets are less commonly addressed. The last column reveals that none of the reviewed approaches explicitly deals with time-oriented data.

**Contribution 0.1 (Time Dependence)** *A concept that explicitly considers the time-dependence of the data for the purpose of feature selection.*

Our approach does not only address the challenges originating from the time-oriented data, but actually leverages the information contained in the time-dependency for the purpose of feature selection. This contribution is opposed to considering all time points as independent. Our method is tuned to continuous and non-cyclic time-dependent data.

**Contribution 0.2 (Transfer of Concepts)** *An application of concepts from an event-based analysis of electronic health records to any numerical data.*

We transfer well-established concepts from the medical domain, where time-oriented data are often event-based, to numerical data from any do-

main. In particular, the idea of aligning temporal data of multiple patients according to a clinical action (e.g. oxygen supply) or the outbreak of a disease (e.g. a cardiac infarction) has been adapted.

**Contribution 0.3 (Model-Free)** *A methodology that does not require the specification of an analytical model of the predictor in advance.*

Our approach does not make assumptions about the type of relationships (e.g. linear, quadratic). It is independent of the future model and can be applied to regression problems where the model class has not been determined yet. This restriction limits the choice of quality measures for candidate feature subsets to generic statistical measures.

**Contribution 0.4 (Feature Subset Evaluation)** *A synchronization approach to evaluate the explanatory power of a feature subset with respect to a target.*

The target's dependence on individual configurations of the subset is evaluated by aligning the different target behaviors that arose prior to and after a configuration. To evaluate the entire feature subset's explanatory power, we provide an overview visualization, which enables analysts to assess the synchronization results across the entire feature space. To date, this overview is only capable of depicting 2D feature subsets.

**Contribution 0.5 (Interactive Visual Analysis System)** *A system of coordinated views and brushing tools for exploration of high-dimensional data.*

The system allows for an efficient exploration of different perspectives, steered by analysis tasks and the current findings in the Synchronization Grid. It can be used to identify a third influencing feature that improves the predictive power when added to the feature subset.

**Contribution 0.6 (Domain Knowledge)** *The possibility to integrate the analyst's expertise into the analysis process.*

Analysts are given the authority to choose the initial features based on their domain knowledge. This could be features that are only influenced by factors outside of the investigated system. Analysts are also guided in deciding on further features to be added based on their analysis findings.

**Contribution 0.7 (Evaluation)** *An application of the approach in the context of vehicle dynamics as well as an analysis of an artificial data set.*

The benefits and limitations of our approach are examined by applying it to real-world time-oriented sensor data, where domain knowledge can be considered, as well as to an artificial data set, where findings can be compared to intentionally included relationships as a ground truth.

# 5 VISUAL FEATURE SELECTION IN TIME SERIES DATA

The overall objective is the modeling of a time-dependent target feature $Y(t)$ using time-dependent predictors $X_1(t), ..., X_n(t)$, where $n < m$ and $m$ is the number of features available in the given data set. The analysis task to be solved is the following: a combination of which time series $X_1(t), ..., X_n(t)$ can replace the information of $Y$? In other words, we have to find those independent features that make up the minimal descriptive feature subset with sufficient explanatory power.

Presenting relationships between independent features and a target in a visual form can support analysts in evaluating the explanatory power of both individual features and feature subsets. Allowing analysts to interact with the visual representations furthermore enables them to assess the influence of individual features on a feature subset's explanatory power. In this sense, Interactive Visual Analysis can help to derive the best-suited feature subset out of the $2^m - 1$ possible candidates.

## 5.1 FEATURE SELECTION WORK FLOW

As an introduction, we will describe a general analysis work flow, which we derived from our understanding of feature selection in the context of time-oriented data. We are not concerned with presenting a full-featured list of steps that should be completed one after the other. The analysis of relationships is often exploratory and cannot be squeezed into a one-branched work flow. Instead, we intend to convey an impression of which aspects need to be addressed on the way. From this overview of a potential work flow, we will derive a number of more concrete requirements.

1. *Choose a target feature.*

   The target is the time series to be modeled. A task involving multiple targets can be reduced to subproblems with only one target.

2. *Choose a number of independent features as initial feature subset.*

   All features that are entirely independent of the investigated system *must* be included in the feature subset, as they capture any target characteristics that cannot be explained by features from the system.

3. *Investigate the relationship between* $Y(t)$ *and* $\{X_1(t), X_2(t)\}$ *independent of time. Are* $X_1(t)$ *and* $X_2(t)$ *sufficient to predict* $Y(t)$*?*

   In general, we distinguish between (1) a relationship affecting the values of $Y(t)$ at a particular time point and (2) a relationship affecting the behavior of $Y(t)$, i.e. from a certain point in time onwards.

   Before explicitly considering the time-related nature of the data, the experts start by evaluating whether the values of $Y(t_0)$ at a certain point $t_0$ in time depend on the values of features $X_1(t_0)$ and $X_2(t_0)$ at the same time point. The temporal order of the given data items is ignored and the data items are regarded as independent.

   In the following, we will present some of the questions that might arise at this stage and clarify their meaning for the analysis.

   (I) *Question:* What does the distribution of the target $Y(t)$ w.r.t. $X_1(t)$ and $X_2(t)$ look like? Can the parameter space $X_1(t) \times X_2(t)$, be divided into regions of similar values for $Y(t)$?

   *Explanation:* If yes, this might indicate a relationship. $X_1(t)$ and $X_2(t)$ could already be sufficient for explaining the target. If this hypothesis is verified, feature selection can be finished.

   If the distribution of target values does not exhibit patterns indicating a relationship, there might be additional features influencing $Y(t)$. The analysis could be continued with step 5.

   (II) *Question:* Can the relevance of $X_1(t)$ and $X_2(t)$ be quantified?

   *Explanation:* If statistical measures indicate that both features contain enough information about the target, the expert might conclude that $X_1(t)$ and $X_2(t)$ explain $Y$ sufficiently well. Such a conclusion should be tested for statistical significance.

   (III) *Question:* Does $Y(t)$ contain values considered as outliers?

   *Explanation:* Outliers might not represent the major characteristics of data and thus distort the regression model. Issues related to outliers are, however, beyond the scope of this thesis.

   (IV) *Question:* Is the distribution of the target feature continuous with respect to the parameter space $X_1(t) \times X_2(t)$?

   *Explanation:* We call a distribution continuous, when the function values do not change arbitrarily within a neighborhood. For example, a high gear engaged in a car is usually associated with high velocity. If those two quantities were independent, engaging the next gear could result in an arbitrary change of

the car's velocity. As there actually is a relationship between those two, slight changes in one quantity result in only slight changes in the other one. To conclude, if the distribution of the target feature is not continuous, this might indicate that $Y(t)$ does not depend on $X_1(t)$ and/or $X_2(t)$.

(V) *Question:* Are there value combinations that do not exist?

*Explanation:* Value combinations that are not available, e.g. due to physical constraints or because they were not caught during sampling, might be special cases to be treated accordingly.

4. *Examine the time-dependence of* $Y(t)$. *How do the values of* $Y(t)$ *develop over time based on a particular value combination of* $X_1(t)$ *and* $X_2(t)$*?*

One objective is to determine whether a combination $(X_1(t_0), X_2(t_0))$ at some time point $t_0$ is a useful predictor for the behavior of $Y(t)$ in a time interval $[t_0, t_i]$ of *reasonable* length $l = t_i - t_0$.

(I) *Question:* How do the values of the target evolve over time?

*Explanation:* Recurring patterns or trends are worth analyzing in more detail. They are in particular interesting when revealing a relationship to the independent features $X_1(t)$ and $X_2(t)$.

(II) *Question:* Do similar value combinations of $X_1(t)$ and $X_2(t)$ yield similar behavior of the target within the time interval?

*Explanation:* Regardless of the model class, we assume that any valid model outputs similar target behavior when being given the same value combinations of $X_1(t)$ and $X_2(t)$ as input.

If the target behaves similarly from those time points onwards, where $X_1(t)$ and $X_2(t)$ exhibit particular values, this is a necessary, but not sufficient, criterion for them to explain $Y(t)$ *for these specific values*. To validate the target behavior for the remaining parameter space, the analyst continues with step 6.

The temporal development of the target values might also be different based on the same values for $X_1(t)$ and $X_2(t)$. In this case, the analyst concludes that there must be an additional feature relating to these differences in the target behavior.

5. *Analyze how differences in the target behavior manifest in other features to identify an additional predictor to be added to the current feature subset.*

If the data items characterizing the differences can be associated with coherent parts of an independent feature's domain, this feature

might be the searched predictor. If it does not introduce any redundancy to the current feature subset, it can be added to improve the subset's predictive power. A reverse-check should be performed to confirm this hypothesis. Continue the analysis with step 3.

6. *Validate the target behavior for the entire parameter space* $X_1(t) \times X_2(t)$.

   The expert might have optimized the feature subset locally: it only explains the target for particular values sufficiently well. The analysis can only be finished, when the target behavior differences have been reduced to a tolerable amount for the entire parameter space.

7. *Repeat steps 3 to 6 with the current feature subset.*

### 5.1.1 *Derived Requirements*

From the work flow described previously, a number of general requirements relating to the extent of the developed approach can be derived.

**Requirement 0.1 (Domain Knowledge)** *Analysts should be able to incorporate their domain knowledge at any step of the feature selection process.*

The work flow starts with an initial feature subset, which results from the analyst's domain knowledge about features that are entirely independent of the investigated system (step 2). In many cases, such features can be taken from the control parameters of a system. In the end, it should be possible to derive an initial subset based on a starting analysis of distributions and patterns in the given data set. We will not consider such an initialization step for now, but it is performed during evaluation of the proposed approach (see Section 7.1). Aside from choosing the start features, domain knowledge can be useful for various other tasks, e.g. when deciding whether to actually add a feature identified as promising.

**Requirement 0.2 (Time-Dependence)** *The time-dependent nature of the data should be considered explicitly.*

This refers to step 4, where the temporal development of both target feature and independent features does play a role, as opposed to step 3, where data items are regarded as entirely independent with no temporal order. Considering the time-dependence is important to get hold of scenarios where the target's values are determined by developments (of both the target and independent features) that happened at previous time points. As an example, this could be an event, e.g. a peak, in one of the independent features leading to a delayed effect in the target feature.

**Requirement 0.3 (Relationship Assessment)** *Analysts should be enabled to identify correlations between independent and dependent features.*

Feature selection thrives on the analysis of relationships to identify those features that contain valuable information about the target feature. Therefore, a primary requirement is the possibility to identify correlations and, in an ideal case, even cause-and-effect relationships. For this purpose, the analyst needs to be able to distinguish between independent and dependent features as well as to assess the relationships between different combinations of those features.

**Requirement 0.4 (Overview and Detail)** *Analysts should be offered both overview and detail visualizations to address the complexity of the data.*

A visual exploration of relationship structures and remaining variance benefits from simultaneously investigating different perspectives on the data. An overview and detail strategy allows for an efficient drill down to interesting relations and for their examination. F+C techniques can also prove useful, in particular for comparing how the target's variance manifests in different independent features (step 5). An IVA system should consequently implement appropriate techniques.

**Requirement 0.5 (Abstraction and Scalability)** *The system should provide a trade-off between scalability and meaningful data representation.*

To enable analysts to focus on the interpretation of depicted patterns, data should be represented with minimal abstraction and information loss where possible. Complex visualizations requiring a high cognitive effort to be interpreted, e.g. projections, should be avoided. Nevertheless, the scalability aspect should not be ignored. Scalability refers to the number of investigated time points as well as the number of features. This relates to the issue of statistical significance. A certain number of data items is necessary to support a relationship hypothesis, otherwise we cannot know whether it simply originated from coincidence. To conclude, aggregations and data abstractions should be used where needed, but kept to a minimum.

**Requirement 0.6 (Model-Free)** *The approach should not prefer certain types of data or relations.*

Not preferring or deferring patterns keeps the approach receptive to all opportunities that might generate a subset with the best quality. To meet this requirement, the approach should not rely on a particular class of the analytical model. This restriction requires the statistical measures that are used to quantify relationships to be highly generic.

## 5.2 THE SYNCHRONIZATION APPROACH

In this section, we describe our synchronization approach to visual assessment of a feature subset's explanatory power. We are not concerned with a conventional regression problem, where the value of the target feature at a time point $t_0$ is modeled by the values of independent features at the same time point. Instead, we search for independent features containing information about the behavior of the target feature, i.e. we do consider the target's behavior within a time interval $[t_0, t_i]$. This means to explicitly make use of information that can be gained from the time-dependence of the data. Synchronization is needed to investigate how certain values of the independent features, which might arise at different time points, relate to the target feature's behavior.

### 5.2.1 *Illustrative Example from the Health Care Domain*

The idea is to synchronize multiple time series, which represent the same target feature, towards a certain event. This synchronization creates the basis for comparing the time series with regard to what happened before and after the event. In this way, the analyst can figure out whether the investigated event can be associated with a certain target behavior. For sequences of temporal events, such a procedure is commonly performed in the health care domain to gain insights about potential precursors and consequences of disease outbreaks. As an illustrative example, consider multiple patients (time series) that are compared regarding how a medical parameter (target) developed shortly before they suffered from a heart attack (event). If, for all patients, the medical parameter strongly increased prior to the heart attack, this indicates a relationship between the medical parameter and the heart attack.

We adapt this event-based concept to numerical time series. An event is then characterized by a value combination of the independent features contained in the current feature subset. The previously considered heart attack as event is replaced by a combination like $\{\texttt{blood oxygen} = 80\%, \texttt{heart rate} = 60\ \texttt{bpm}\}$. Instead of multiple patients, we now consider only one patient, for whom we investigate whether the event can be related to the target feature $\texttt{blood pressure}$. The event $\{\texttt{blood oxygen} = 80\%, \texttt{heart rate} = 60\ \texttt{bpm}\}$ can arise at multiple time points throughout the captured time interval. The time series to be synchronized are then actually sections of the patient's $\texttt{blood pressure}$ development. These time series sections are chosen in such a way that the center of the cov-

ered time interval is one of the time points, at which the investigated event occurred. In this way, we can compare the blood pressure behaviors that the same patient exhibited each time the event occurred. If the behavior exhibits large differences based on the same event, we can conclude that there is no explicit relationship, thus indicating that blood oxygen and heart rate do not contain sufficient information to explain blood pressure. This kind of synchronization is similar to comparing the target behavior starting every Monday (event), when considering a cyclic time series containing daily values per week.

### 5.2.2 *Fundamentals*

The feature selection is the first step of statistical modeling and entirely independent on the following model building step. No explicit regression model – in terms of model type or parameters – is given at any stage of the procedure. Therefore, we treat the model associated with a candidate feature subset as a *black box model*, which is solely observed in terms of its inputs and outputs. For this reason, we put the explanatory power of the underlying feature subset on a level with the quality of the black box model, which are both to be optimized. In each iteration of the feature selection procedure, there are two main questions: (1) Is the explanatory power of the feature subset sufficient to explain the target feature? (2) If not, which feature do I need to add to increase the explanatory power?

A black box model makes it challenging to quantify its quality – and therefore the explanatory power of the underlying feature subset –, because we cannot simply compute a model error. But even without a concrete specification of the model, general assumptions about the regression problem can be made, which are useful for an evaluation of a feature subset's explanatory power and thus the quality of a future model.

**Assumption 0.1 (Prediction)** *If we consider any valid regression model, we can assume that it outputs equal predictions based on the same input, independent of the input's time of occurrence.*

This assumption is derived from the fact that a regression model is nothing more than a function $f$ with domain $X$, for which the following applies: $\forall a, b \in X : a = b \implies f(a) = f(b)$. It refers back to step 4 of the feature selection work flow. The negation of this implication can be described by: $\forall a, b \in X : f(a)! = f(b) \implies a! = b$. This, in turn, implies: if $f(a)! = f(b) \wedge a = b$, then $f$ is incomplete. Thus, looking at assumption 0.1 the opposite way, we get the following statement:

**Assumption 0.2 (Remaining Variance)** *If the model outputs different predictions for the same input, this means that the underlying feature subset is not sufficient to explain the target feature.*

This conclusion is the fundamental argument, upon which the entire synchronization approach is built.

### 5.2.3 *The Leading Idea*

How can these assumptions about a regression problem help us to determine the explanatory power of a feature subset? A feature subset sufficiently explains the target feature, if the associated black box model predicts the target values for new data with the property specified in assumption 0.1. With the derived assumption 0.2 in mind, we relate the explanatory power of a feature subset to the differences in predicted target values. Following the statistical definition of variance as a measure of a data set's spread, we refer to these differences as *target variance*. Consequently, the target variance is a measure for how much the predicted target values differ when the model receives the same input. According to assumption 0.2, it is an indicator for those portions of the target feature that a feature subset cannot explain.

**Assumption 0.3 (Explanatory Power)** *A feature subset has a high explanatory power, if the target variance is low, i.e. if the target behaves similarly starting from the same input configurations.*

To evaluate a feature subset, training data is required. Equal inputs are defined as data items that exhibit the same values for those features contained in the feature subset. We refer to this as a *configuration*. As an example, a configuration could be $\{\mathtt{velocity} = 50\ \mathrm{km/h}, \mathtt{road\ curvature} = 0.3\}$. The corresponding values of the target feature are referred to as the predictions. As we deal with time-oriented data, we intend to explicitly make use of the data's time-dependence to achieve more stable and meaningful results. For this reason, we do not consider scalar values as predictions, but investigate whether the input relates to a certain target behavior that arises temporally close to the occurrence of the input configuration.

#### SIMPLIFIED ILLUSTRATIVE EXAMPLE

How does the visual assessment of the target variance for a given feature subset work? That is, how can we determine whether it is indicating an acceptable explanatory power of the feature subset? To explain the working principle of our synchronization approach, we consider a simplified

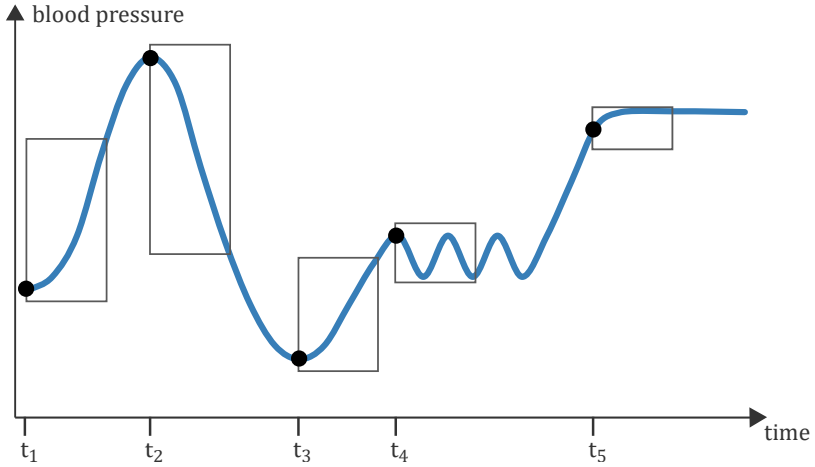



**Figure 15:** *A fictional time series depicting* blood pressure (BP) *as a function of time. Considering another feature* heart rate (HR)*, we investigate whether* BP *depends on* HR *rather than on* t *alone. We compare the behavior of* BP *from synchronization points* $t_i$ *with* $HR(t_i) = 100$ bpm *onwards, within a certain interval (rectangles).*

illustrative example. Given a feature subset containing only one feature heart rate, we specify a value of 100 bpm as initial configuration. This value can arise at multiple points $t_i$ throughout the captured time range. We are now interested in whether blood pressure behaves similarly each time the heart rate reaches 100 bpm or if the configuration is followed by different behaviors. According to assumption 0.3, this allows us to determine whether heart rate might be a useful predictor for the behavior of blood pressure in a time interval $[t_i, t_{i+1}]$ (see step 4 of the work flow).

For this purpose, we compare the developments of blood pressure values from each $t_i$ onwards (Figure 15). The $t_i$ are called *synchronization points*. By exhibiting the same configuration heart rate $= 100$ bpm, they mark the common baseline for comparison. We investigate the behavior of blood pressure within a *synchronization interval* starting from each $t_i$ (Figure 15, rectangles). In an ordinary time series plot like in Figure 15, the synchronization points are distributed along the time axis. Consequently, comparing the corresponding time series sections within the synchronization intervals is challenging and requires a lot of cognitive effort, because a common baseline is missing. This prevents analysts from efficiently perceiving differences among the curve sections and does not support an evaluation of the target variance.

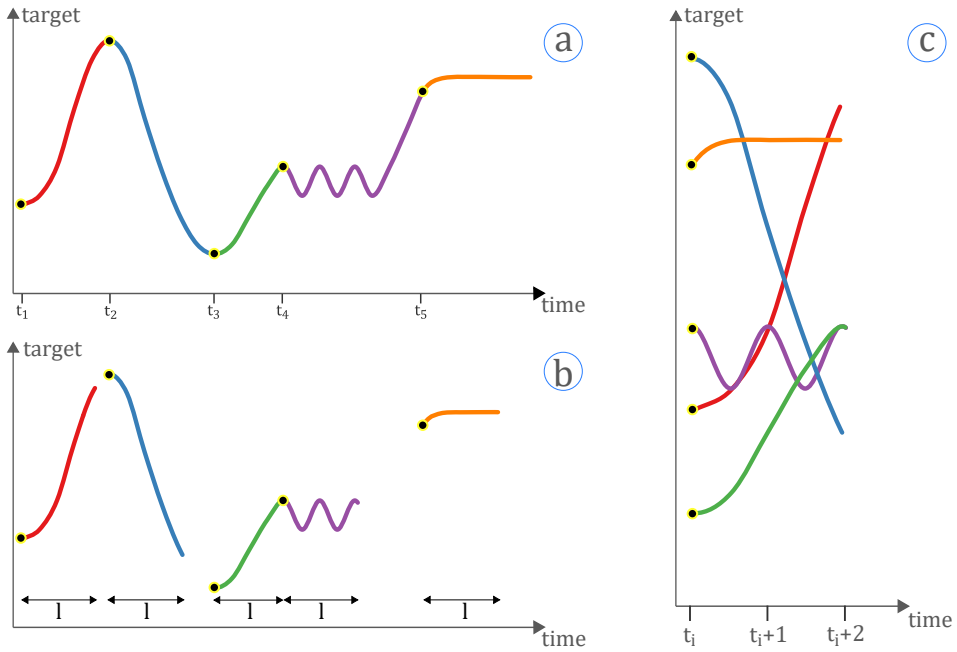



**Figure 16:** *Synchronization of target behavior towards identical configurations: the synchronization points are used as split points for the target time series (a). The resulting curve sections are cropped to synchronization interval length* $l$ *(b) and shifted along the time axis (c).*

### SYNCHRONIZATION OF TIME SERIES SECTIONS

To simplify the comparison, we suggest to horizontally align the synchronization points $t_i$, together with the time series sections bound to them. We call the temporal alignment of identical configurations *synchronization*, because it allows the analyst to directly compare the different temporal developments originating from the same configuration. This synchronization is performed in three steps: (1) the original target time series is split at the synchronization points, (2) the resulting curve sections are cropped to a certain interval length, and (3) are shifted along the x-axis, such that the associated synchronization points share the same x-coordinate – the zero line, to be more precise. These three steps are shown in Figure 16. As a result, all curve sections have a joint reference point (i.e. the synchronization points) as well as equally long observation periods (i.e. the synchronization intervals) and can be compared directly.

After synchronization, analysts can directly compare the temporal developments of $\mathtt{blood\ pressure}$ that originate from the same configuration $\mathtt{heart\ rate} = 100\ \mathtt{bpm}$ with little cognitive effort. Using their visual per-

ception and human sense of curve similarity, they can efficiently evaluate the presence of target variance by simply "seeing" how widely the time series sections are spread. This frees them from the need of defining similarity as a measure at this early stage, which would require the specification of a model. If the curves exhibit similar shapes, the variance is low, as opposed to curves that clearly diverge. Following assumption 0.3, low variance indicates a relation between $\mathtt{heart\ rate} = 100\ \mathtt{bpm}$ and the development of $\mathtt{blood\ pressure}$ shortly afterwards. In contrast to that, diverging curves raise the question of one or more additional features that influence $\mathtt{blood\ pressure}$ to behave either this or that way.

Uncertainty emerges from the naïve assumption that the heart rate of $100\ \mathtt{bpm}$ does not change within the synchronization interval. If the values of $\mathtt{blood\ pressure}$ change, although $\mathtt{heart\ rate}$ is assumed to stay the same, this is unexpected behavior according to Q3 in Table 1 and can be explained by additional influencing factors. However, in general, we cannot assure that the heart rate stays the same throughout the synchronization interval. This makes it challenging to distinguish between changes in $\mathtt{blood\ pressure}$ that are related to additional influencing factors and changes that are caused by (unexpected) changes in $\mathtt{heart\ rate}$ and thus can be categorized as expected behavior according to Q1 in Table 1.

### 5.2.4 *Synchronization Line Plot*

The visualization depicting the results of a synchronization is called *Synchronization Line Plot* (Figure 16c). Each curve is associated with exactly one synchronization point, from which onwards it depicts the target's temporal development within a certain time interval. Contrary to the traditional Time Series Plot, the x-axis does not represent absolute time values. Instead, it represents the number of seconds that have passed since the respective synchronization point was reached. The mapping between the absolute time $t_{abs}$ and the relative time $t_{rel}$ is defined as $t_{abs} = t_i + t_{rel}$, where $t_i$ is the synchronization point.

Up to now, we only consider the time points that *follow* the synchronization point. However, there might be cases, where it is beneficial to examine how the target values evolved towards the synchronization point. This requires to additionally display the time points *preceding* the synchronization point. Different appearances of the Synchronization Line Plot are shown in Figure 17. Observing the target behavior both before and after the synchronization point might provide additional insights, for exam-

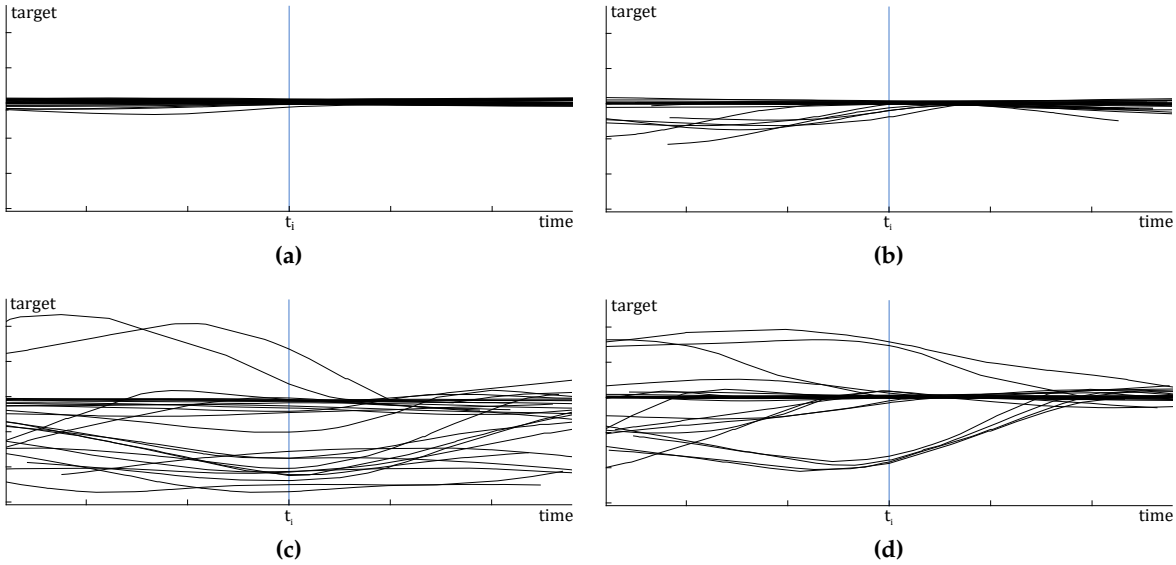



**Figure 17:** *Synchronization Line Plots depicting the target behavior around 26-35 synchronization points $t_i$: no target variance (a), low target variance (b), high target variance (c), and high variance arising mainly before the synchronization point (d).*

ple when a certain variance can be observed before the synchronization point, but the curves behave similarly afterwards. However, the meaning of such a pattern for the relationship between independent features and the target is still to be investigated.

Visually assessing the target variance works well for a reasonable number of displayed curves in the Synchronization Line Plot. However, there might be cases where the same configuration arises at a large number of time points, resulting in a correspondingly large number of curves. In this case, visual clutter might prevent the analyst from deriving reasonable conclusions. Semi-transparent curves, which add up to more saturated colors at overlaps, have already been proposed to provide insight into the individual course of curves, even though many of them are depicted at once [58]. We apply a non-linear, strictly decreasing transformation function, which maps the number of curves to an opacity value in $[0, 1]$.

## 5.3 THE SYNCHRONIZATION GRID

In the previous sections, we described how synchronization can be used to perceive and analyze the target variance, when being given a configuration like $\mathtt{heart\ rate} = 100\ \mathtt{bpm}$. Note that all conclusions, which were drawn up to now, are based on this particular configuration. They do not provide any information about the target variance when considering another configuration like $\mathtt{heart\ rate} = 60\ \mathtt{bpm}$. Consequently, a conclusion about the explanatory power of a feature subset cannot be confirmed or rejected, until the target variance has been assessed for the entire feature space. We need to check the target variance *for each possible configuration*.

As our approach is not tuned to medical data sets, but targeted at multivariate time-oriented data in general, we will abstract the illustrative example in the following. Let us consider two independent features $X_1(t)$ and $X_2(t)$ with configurations $\{x_1, x_2\}$. A configuration corresponds to one single point in the two-dimensional feature space $X_1(t) \times X_2(t)$. Aiming at an analysis of the entire feature space, we run into a scalability problem: the number of available configurations for two numerical features might be huge, because two data items are unlikely to exhibit exactly the same decimal numbers. When no two data items are equal, the number of required Synchronization Line Plots (one per configuration) equals the number of data items. Viewing and analyzing ten thousands of plots to cover the entire feature space is infeasible. Another problem lies in one configuration being associated with exactly one time point, which results in each plot containing only one curve. Intuitively, this makes it impossible to determine the target variance.

### 5.3.1 *Grouping of Configurations*

To start an analysis with a space-efficient overview of the target variance across the entire feature space, we systematically group the available configurations, instead of investigating each configuration individually. Each group is then visualized by one Synchronization Line Plot, thus reducing the number of required plots and increasing the number of curves per plot at the same time. Grouping softens the conditions in our assumptions 0.1 to 0.3 from considering the *same* inputs to *similar* inputs. Thus, configurations in the same group should be highly similar in terms of their distance in the feature space. Each group might be represented by a *representative configuration*, similar to the centroid in k-means clustering.

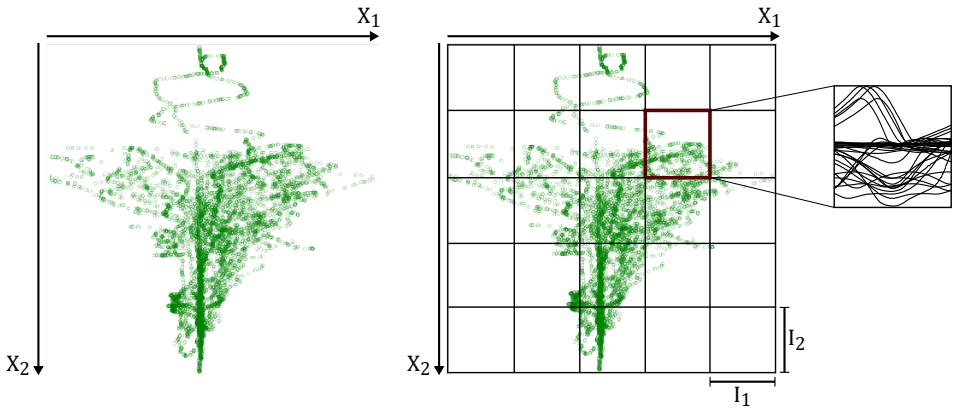



**Figure 18:** *Grouping of configurations: a scatter plot depicts the configuration space (left). A grid overlay indicates the grouping strategy: configurations in the same cell belong to one group (right).*

Let's stay with the example of two independent features $X_1(t)$ and $X_2(t)$. The configurations $\{x_1, x_2\}$ within the feature space can be depicted as points in a scatter plot (Figure 18, left). The most straightforward solution to grouping these configurations is a binning strategy that is applied to both features and divides the two-dimensional feature space into a number of *cells*. This can be imagined as an axis-aligned grid that is laid on top of the scatter plot (Figure 18, right). The resolution in both dimensions can be arbitrarily chosen to meet the analyst's needs. Consequences of choosing different resolutions are discussed in Section 5.4.

As a result of binning, a cell is defined by two intervals: $I_1$ for the $X_1(t)$ and $I_2$ for the $X_2(t)$ dimension. Each configuration can then be assigned to a cell, such that its values $x_1$ for $X_1(t)$ and $x_2$ for $X_2(t)$ are contained in the cell's respective intervals $I_1$ and $I_2$. Some cells might stay empty during the assignment of configurations. They originate from empty regions in the configuration space, where the corresponding configurations do not exist in the data set. This might be due to physical constraints, e.g. high road curvature and high velocity are unlikely to occur as a configuration. The result of grouping is a grid representing the feature space, where configurations in the same cell belong to one group. Each non-empty group of configurations can be seen as a set of similar inputs – approximated by a representative configuration –, for which the target variance is visualized using a Synchronization Line Plot (Figure 18, right).

### 5.3.2 *Overview Visualization*

In Section 5.2.4, we have introduced the Synchronization Line Plot as a visualization depicting the target variance emerging from one specific configuration. Section 5.3.1 addresses an overview of the configuration space that consists of cells, each of which holds a number of configurations. To achieve an overview visualization depicting the target variance across the entire feature space, we put these two techniques together.

We perform the synchronization for each configuration group using the time points that are associated with its members. The resulting time series sections depict the temporal development of the target feature based on the group's representative configuration. We then visualize the curves belonging to each group using a Synchronization Line Plot. The individual Synchronization Line Plots representing each group of configurations are combined to an overview visualization. To ease the interpretation of the overview visualization, the spatial relation between a Synchronization Line Plot and the depicted portion of the feature space should be obvious from the arrangement of plots. It is therefore intuitive to again apply the grid layout, which was already used for grouping (Figure 19). Consequently, we name the resulting visualization *Synchronization Grid*.

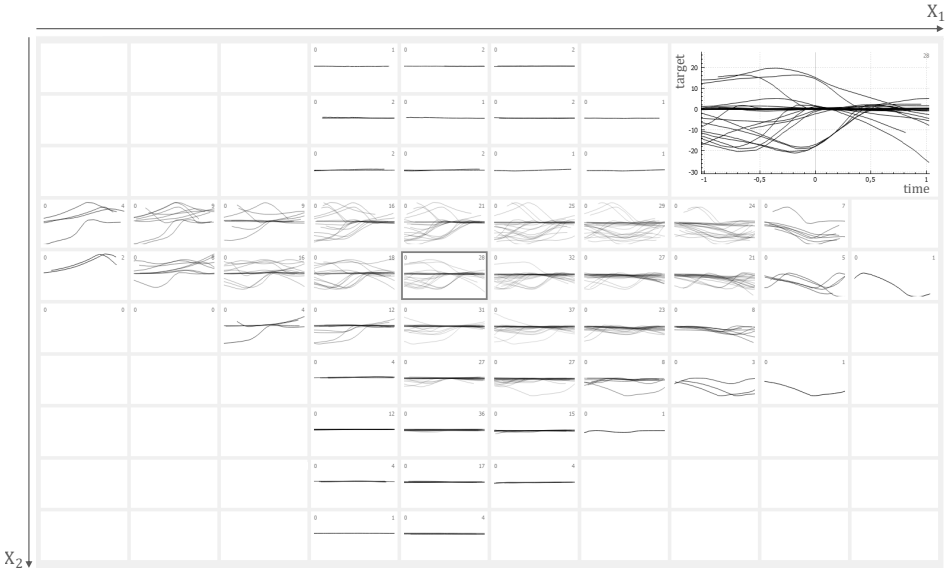



**Figure 19:** *Synchronization Grid: visualizing the target variance across the entire feature space* $X_1(t) \times X_2(t)$. *Synchronization Line Plots depict the variance for groups of similar configurations (cells). An enlarged focus plot allows for detailed investigation and interaction (top right).*

The resolution of the grid and therefore the maximum space for one cell determine the size of the Synchronization Line Plots. When the screen space is allocated to a large number of plots, it is important to fully use the available space of an individual plot. As we do not aim at a quantification of the target variance, but simply rely on a visual assessment, it is sufficient to observe the spread of the curves relative to each other. To reduce visual clutter to a minimum, we omit axes and labels and use Small Multiples as a visualization technique to display the synchronized curves.

Arranging the Synchronization Line Plots side by side allows for a direct comparison of the target variances for different representative configurations. Being able to view the target variance across the entire feature space at once, the analyst can efficiently evaluate the explanatory power of the feature subset. Critical cells exhibiting a large target variance, which is not desired, can be identified with little effort. The identification of critical parts is not restricted to individual cells: analysts are also enabled to observe where large target variance spreads across larger parts of the feature space. This leads back to the inter-dependencies between feature subset and item subset, whose investigation is not part of this thesis.

### 5.3.3 *Focus Plot*

In case the Synchronization Grid reveals critical cells, the next step towards feature subset refinement is to carefully examine the corresponding target variance in more detail. To steer the analysis towards regions of interest, we realize a Focus + Context approach: hovering one of the Synchronization Line Plots moves it to focus, while keeping the remaining plots, which depict the variance for the entire feature space, as context. Moving a plot to focus means to enlarge it for a detailed analysis of the depicted curves. Axes and labels are added to provide an orientation for the assessment of interval length (time axis) and actual value range of the plotted target feature (y-axis). Standard interaction techniques like zooming and panning enable the analyst to take a closer look at subsets and subsections of curves, which is especially useful when examining a large number of curves. Such an exploration might result in a subset of curves that the analyst intends to further investigate. Taking advantage of the concept of Coordinated Multiple Views, the analyst can brush these curves in the Focus Plot to analyze their equivalents in other views. Brushing and linking within the Synchronization Grid as well as the linking to other visualizations are addressed in detail in Section 5.5.

Enlarging a plot at its position would occlude the neighboring plots and therefore a part of the context visualization. We need to adjust the placement of the *Focus Plot*, such that it does not impair the perception of the context. In an ideal case, the focus plot can be integrated into the Synchronization Grid to allow a simultaneous analysis of focus and context information that does not require interaction like switching windows. Depending on the distribution of configurations within the feature space, empty cells might cover a larger contiguous area. Those cells occupy space in the Synchronization Grid, although they do not convey any information. We can use this unexploited space to display the enlarged Focus Plot. For this purpose, we solve a simplified variant of the maximum-area empty rectangle problem [59] to identify the largest, axis-aligned, horizontally oriented rectangular block of empty cells in the grid (Figure 19, top right).

## 5.4 DESIGN DECISIONS

Along the way of building up the Synchronization Grid, several design decisions have to be made. Those with the largest impact on the visualization and interpretation of the data are discussed here: (1) the length of the synchronization interval, (2) the resolution of the grid that is used for grouping, and (3) those configurations of each cell that are actually taken for synchronization. We will see in the following that all of these aspects are actually inter-dependent.

### 5.4.1 *Length of Synchronization Interval*

The synchronization interval is the time interval, in which the target behaviors are compared to determine the variance. It contains the synchronization time point, either at its left border or center, depending on whether the behaviors are investigated from the synchronization point onwards or both before and after this point. The question is: which length is reasonable for a synchronization interval?

#### CONSTANT OR ADAPTIVE INTERVAL LENGTH?

Before thinking about the interval length, we need to decide whether it should be constant or adaptive. A *constant length* means that the same number of time points preceding and following the synchronization point is considered. As opposed to that, one could take only time points around the synchronization point, for which the values of the initial features change within a certain tolerance. We refer to this as *adaptive length*.

With a constant length, the synchronization interval might cover changes in the initial features. In this case, the naïve assumption of unaffected independent features does not hold. As a consequence, we cannot not know whether target variance arises due to an additional influence measure or simply as an expected result of changes in the initial features (see Table 1, Q1). The advantage of covering such changes lies in the possibility to investigate the derivatives of initial features as potential predictors. When deciding for an adaptive interval length, the opposite holds: it has no interference with changing initial features, but does not include derivatives.

For the decision for a constant or adaptive interval, we also need to consider another fundamental issue: the comparability of time series sections. It can be approached at two levels: (1) within the same Synchronization Line Plot and (2) within the Synchronization Grid. The subject of the first is whether all curves referring to the same group of configurations are depicted over the same range on the time axis. The second refers to how comparability could be preserved across multiple Synchronization Line Plots by sharing interval lengths, e.g. along a column of the grid.

The assessment of the target variance entirely builds upon the perceived similarity of curves. To be able to define a notion of similarity, it is highly important that those curves are comparable. Where the lengths of curves are varying, i.e. for non-constant interval lengths, some curve sections do not have a counterpart for comparison. Thus, information necessary for similarity assessment are missing. For this reason, we decide for a *constant interval length* within one Synchronization Line Plot, such that all curves are depicted across the same time period. Regarding the Synchronization Grid, comparability between cells is difficult to achieve, because the curves are depicted in varying contexts, i.e. regarding different representative configurations. Independent of the interval length being constant or not, curves from different cells are thus not comparable. Nevertheless, for the purpose of consistency, we choose the same constant interval length for the entire Synchronization Grid.

#### CHOOSING THE ACTUAL CONSTANT INTERVAL LENGTH

The next step is to specify the actual length for the synchronization interval. If the interval is too small, patterns in the target's temporal behavior do not have time to evolve. In an extreme case, the interval does not contain enough time points to reliably consider the temporal development of values for analysis, which leaves us with an approximation of the relationship between feature values at the same time point (step 3). On the

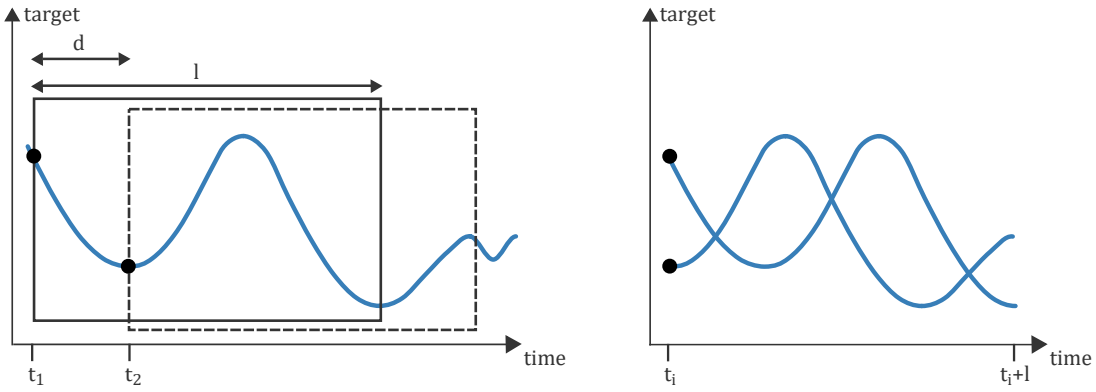



**Figure 20:** *Undesired replication of curve sections as a result of the interval length* l *being larger than the distance* d *between synchronization points. Such redundancy does not convey valuable information.*

other hand, if we choose the length too large, the informative value of existing target variance decreases, because we likely include a time point, at which the synchronized curves behave different. However, such dissimilarity is not meaningful. A large interval might also disprove the naïve assumption of an unvarying configuration within the entire interval, as the configuration has more time to change.

The choice of a suitable interval length l also depends on the distance d of the synchronization time points. If l is significantly larger than d, the tail end of the interval overlaps the beginning of the following synchronization point's interval (Figure 20, left). In the Synchronization Line Plot, the curve section contained in the overlap appears as a *replication*, because it is covered by both intervals (Figure 20, right). Due to their redundancy, a comparison of the replicated curves corresponding to the two synchronization points does not provide valuable information.

### 5.4.2 *Resolution of Aggregation Grid*

Before we can perform the actual synchronization, we need to define the grid's resolution. An individual cell represents a configuration with a certain tolerance. An imaginary representation of this representative configuration could be the center of the cell. The grid resolution has an impact on different quantities: (1) the size of the Synchronization Line Plots, (2) the scope of a configuration represented by a cell, and (3) the level of detail that is conveyed by the visualization.

The vagueness of the representative configuration associated with a cell increases with decreasing resolution: the larger a cell, the larger the spread of the configurations describing the representative configuration. At the same time, larger cells tend to contain a larger number of configurations, resulting in more curves being depicted in the Synchronization Line Plot. This increases the statistical significance of derived conclusions. However, the number of synchronization points cannot be increased arbitrarily to achieve statistical significance due to issues like the replication problem.

#### PARTITIONING STRATEGY

For subdividing the feature space into a number of cells, we make use of a domain-uniform partitioning: space represents uniform partitions of the domains between the respective minimum and maximum values for both features. For further information, Mühlbacher and Piringer provide a discussion of different combinations of domain-uniform and frequency-uniform layouts [57]. The maximal grid resolution is determined by the screen space and the minimum size of a Synchronization Line Plot. Although a plot's content is reduced to the most relevant elements, a minimum size is needed to ensure that analysts can effectively perceive the depicted curves. We empirically determined this minimum size to be 80 pixels for both width and height. Consequently, the grid is not arbitrarily scalable with respect to the number of cells. However, the missing scalability is acceptable at this point, because the Synchronization Grid itself is only thought of as an overview visualization, whose purpose is to present adequate portions of the feature space as a starting point for drill-down. To ease the interpretation of the overview, we start with the same resolution for both directions, such that the initial binning of the feature space is not be skewed. As a grid resolution that balances all of the three aspects named above (plot size, configuration scope, level of detail), we determined a subdivision into 10 cells in each direction. As the analysis proceeds, this resolution can be adjusted to the analysts needs.

### 5.4.3 *Selecting Configurations for Synchronization*

#### MANUAL SELECTION LEADING TO SYNCHRONIZATION GRID

The original idea that led the way to the Synchronization Grid was to let the analyst manually choose interesting configurations towards which the target time series is synchronized. When depicting two-dimensional configurations in a scatter plot, manually choosing a configuration simply means to select a scatter plot item. The remaining items representing the entire configuration space can be viewed as a context during selection.

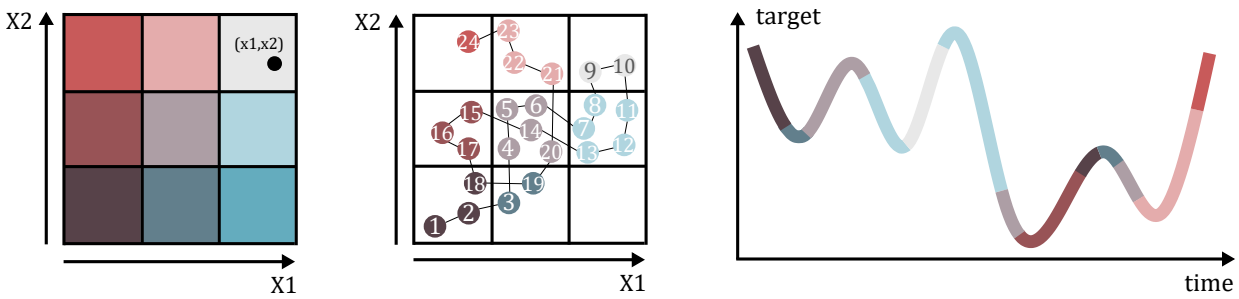



**Figure 21:** *Configuration trajectory: configurations being assigned to cells (left) one by one in temporal order leave a trace through the feature space (center). This assignment is mapped to the target time series (right). Configurations within the same cell are often successive in time.*

We will not again discuss the significance problems arising from each configuration being likely to appear only once in the data set (see Section 5.3). Manually selecting and investigating individual configurations one by one does not support an overview of the synchronization results for different configurations.

To enable the analyst to explore the synchronization results for the entire feature space, a lens-based approach is conceivable. A lens could be implemented as a rectangular area that is moved across the scatter plot and displays the Synchronization Line Plot for the currently covered configuration. However, using such a lens, users can only investigate one plot at a time, while viewing the remaining configuration space in the form of scatter plot items as context. This prevents the analyst from comparing the synchronization results for different configurations and gaining an overview of the configuration space. These problems led us to the idea of statically arranging a number of Synchronization Line Plots next to each other, such that they cover the entire feature space.

#### AUTOMATED FILTERING OF SYNCHRONIZATION POINTS

From the thoughts described above, we developed the Synchronization Grid consisting of Synchronization Line Plots, each of which depicts the target behaviors that resulted from synchronization towards a group of similar configurations. However, the configuration groups as resulted from grouping cannot be used for synchronization right away. During grouping, the configurations are assigned to the cells one by one in the temporal order of their occurrence. In doing so, they leave an imaginary trace in the feature space (Figure 21, center), which we call *configuration*

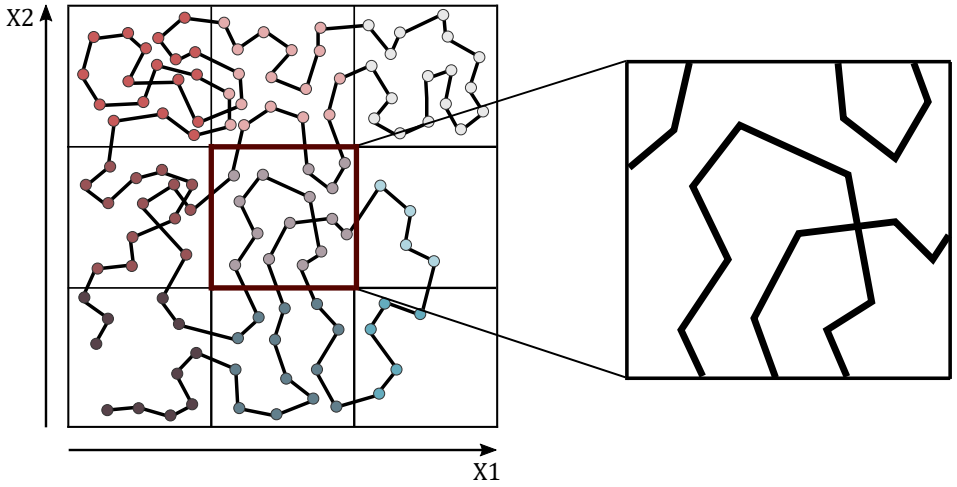
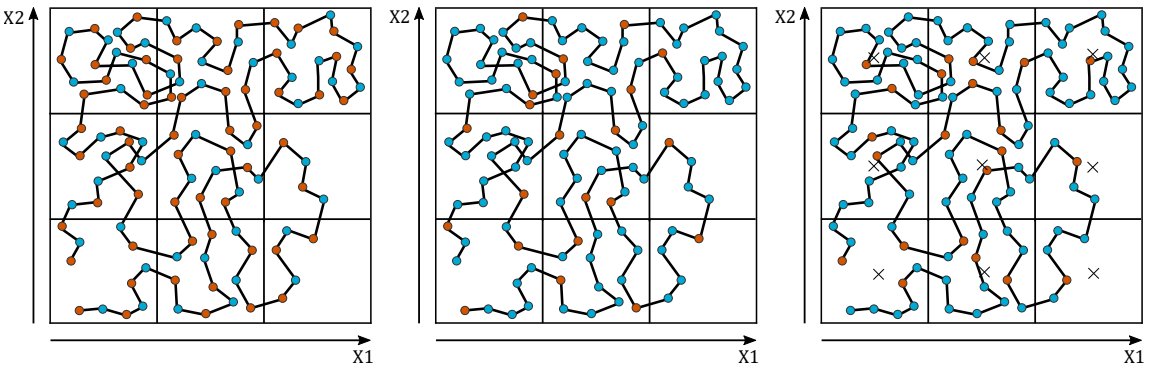



**Figure 22:** *A more complex configuration trajectory of 114 points (left). It enters the middle cell four times, resulting in four trajectory sections (right).*

*trajectory*. As each configuration – and therefore its assigned cell – is associated with a time point, we can map the cell assignment to the target time series (Figure 21, right). As large sections of the time series are actually of the same color, many of the configurations assigned to the same cell are actually directly successive in time. This is due to the state of a stable system not changing arbitrarily, but rather smoothly, over time. If we perform a synchronization using all these successive time points in a cell, this will result in a large number of replicated curves. A comparison then does not provide any information concerning the relation between the target and the independent features.

To avoid such replications, we need to select a subset of the configurations in each cell prior to synchronization. We will discuss different approaches to this selection. Most of them are based on the configuration trajectory that builds up in the feature space during grouping (Figure 22, left). When moving forward in time, this trajectory might cross an individual cell multiple times, resulting in a number of *trajectory sections* (Figure 22, right). We will describe and discuss the selection of synchronization points by an example of one single cell. Basically, we search for a subset of synchronization points that (1) represent the cell's representative configuration well, (2) produce a reasonable number of synchronized curves, and (3) are not too close to each other with respect to time. Synchronization points can be chosen for each cell in three different ways:




**(a)** *Regular Sampling* **(b)** *Section Beginnings* **(c)** *Closest to Reference Point*

**Figure 23:** *Filtering of configurations: different approaches applied to the same configuration trajectory. Orange points mark selected configurations, while those depicted in blue are not considered for synchronization.*

1. *Regular Sampling:* Consider the time points assigned to the cell in their temporal order. Starting at the first time point, sample the points with a sampling interval of size d (Figure 23a).

2. *Beginnings of Trajectory Sections:* Consider the trajectory sections that cross a cell. From each section, take the first time point (Figure 23b).

3. *Closest to Reference Point:* Again, consider the trajectory sections. Additionally, let's define a point within the cell, e.g. its center $c = (x_1, x_2)$, as the representative configuration, which serves as a reference point. From each section crossing the cell, now take the time point that is closest (in terms of Euclidean distance in the feature space) to that reference point (Figure 23c).

Table 4 shows a comparison of their benefits and drawbacks. The main advantage of a regular sampling is that multiple synchronization points are taken per trajectory section, such that long sections are exploited. This results in a better coverage of the feature subspace within the cell, which gives more emphasis to conclusions drawn from the synchronization results. However, an undesired replication of patterns is more likely than for the other two approaches, where only one synchronization point is taken per trajectory section. Apart from replication being less likely, taking the starting points of the trajectory sections does not provide any further advantage, except for the little effort needed for implementation.

| | 1: Regular Sampling | 2: Beginnings of Trajectory Sections | 3: **Closest to Reference Point** |
|---|---|---|---|
| + | Evenly distributed synchronization points | Replication is less likely | Replication is less likely |
| | Long trajectory sections are exploited | Straight-forward to implement | Actual configurations closest to intended configuration |
| | | | Investigated configuration specified flexibly by reference point |
| − | Patterns might be replicated → no valuable information | Long sections are only represented by one synchronization point | Long sections are only represented by one synchronization point |
| | Sample interval needs to be chosen | Synchronization points are distributed irregularly | Synchronization points are distributed irregularly |
| | | | Reference point has to be chosen |

**Table 4:** *Benefits and drawbacks for different approaches to the filtering of synchronization points. Approach 3 provides the best trade-off for the generation and interpretation of Synchronization Line Plots.*

The third approach offers strong additional benefits. Remember that the configurations contained in a cell are viewed as variants of the same representative configuration, which we suggest to imagine as the cell's center. However, by choosing the reference point, the investigated configuration can be flexibly adjusted. The configurations selected from each trajectory section are then the most similar to the representative configuration, thus providing an accurate approximation. The approach might actually be combined with regular sampling to adequately make use of very long trajectory sections. In this case, also points with only locally minimal distances to the reference point are selected from a section. At the same time, the drawbacks of this approach weigh relatively little and might be solvable with little effort. For example, the choice of an initial reference point can be intuitively solved by selecting the cell's center. As a conclusion, we decide for the *Closest to Reference Point* approach, because it offers benefits that cannot be achieved otherwise.

### 5.4.4 *Summary*

Several design decisions had to be made on the way to building our overview visualization. As a result of our examination, we have proposed (1) to use a constant interval length, (2) to apply an initial grid resolution of ten cells in each direction, and (3) to choose the synchronization points according to their distance to a reference point within the cell.

All three visualization parameters need to be specified: the grid resolution, the interval length, and the reference point. Guiding the analyst by providing reasonable initial choices allows her to concentrate on the actual analysis and achieve first results, rather than figuring out suitable parameter values. Currently, naïve choices are made by the system, but allowing the user to adjust these values in the future might offer an additional benefit. The main issue with constant interval length and the reference point approach is that a target variance within a Synchronization Line Plot's can arise from both varying initial features as well as an additional influence measure. Consequently, the analyst has to make an additional effort to distinguish both cases.

At the same time, the constant synchronization interval length does not exclude the possibility to consider changes of the independent features, i.e. first-order derivatives, as potential predictors in the future. Filtering the synchronization points with respect to the reference point ensures that replicated curve sections are less likely to arise, because each trajectory section is only represented by one synchronization point, which favors sufficiently large distances between synchronization points. Most important for an analysis of the target variance is the comparability of curves within a Synchronization Line Plot, which is ensured by the constant interval length, because all curves in a plot are of equal length.

## 5.5 LINKING TO OTHER VIEWS

Different perspectives on the data have to be considered to reveal hidden relationships and patterns. Being able to connect multiple perspectives might offer better insights than considering the visualizations independently [40]. We achieve this by linking the Synchronization Grid to other available views in the system. This concept also serves the overview and detail strategy [71], which has proven useful in various applications.

First of all, the components of the Synchronization grid are connected with one another. User interaction in the Focus Plot leads to an update of the Small Multiples representing the entire feature space as context. This enables the user to assess how patterns propagate through the feature space. The concrete realization is described in Section 5.5.1. As a second step, the Synchronization Grid as a whole is linked to standard visualization techniques like a Time Series Plot, a Dot Plot, and a Histogram. This serves multiple purposes: getting an idea where synchronized curves are located within the global time range and which time points were actually used for synchronization (Section 5.5.2), identifying additional features that influence the target (Section 5.5.3), or investigating details for individual data items where needed (Section 5.5.4).

### 5.5.1 *Brushing in Synchronization Grid*

In general, brushing is nothing more than a selection of data items, which a user performs by interacting with the available visual representations of data. In most cases, such selections are used for drill-down. In this sense, the analysis is steered towards regions of interest, which are defined by analysts based on their visual impressions. Steering works particularly well when composite brushing is supported, where brushes can be combined using Boolean operations, similar to a database query. As brushing is a widespread technique in Information Visualization, various kinds of brushes have been proposed in literature. In the following, we will discuss some brushing mechanisms concerning their suitability for curves as the underlying visual representation.

Brushing primarily serves the purpose of analyzing the target variance in individual cells of the Synchronization Grid. If a cell exhibiting a certain variance is identified, the analyst wants to further investigate the varying curves in the corresponding Synchronization Line Plot. Target variance can arise in different forms, from individual curves strongly deviating from the remaining set over different groups of curves diverging to curves being distributed all over the co-domain. Using her visual sense, the analyst can determine a number of curves that incorporate the majority of the target variance. This subset of curves is brushed as the *region of interest* for further analysis, which often involves the search for an additional feature relating to the incorporated variance.

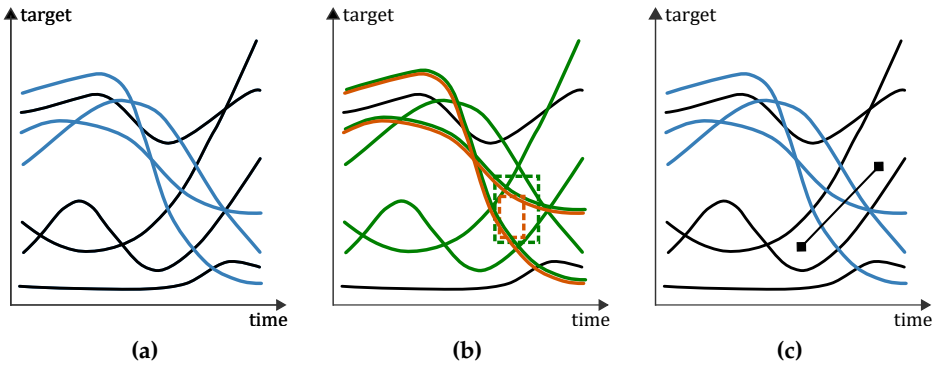



**Figure 24:** *Selecting the blue curves in a dense area (a) is not possible with a rectangular brush. It either selects too few or too many curves (b). The line brush is superior in this case (c).*

### BRUSHING MECHANISMS

Brushing is carried out in the Focus Plot of the Synchronization Grid. The most common kinds of brushes are line brush, rectangular brush, angular brush, and the query-by-example brush. Angular brushes [29] and brushing by a specified target function [33] are not meaningful in our case. Both are based on assumptions about the slope or even shape of the elements to be brushed. In contrast, we need an efficient way of specifying a subset of curves based on the analyst's visual perception of their developments.

The most common technique for selecting an area in various applications is the selection rectangle. It can be found as a tool in most software programs, be it a text editor, an image-editing program, or high-level administration software. As it is often used in daily practice, analysts are highly familiar with this technique. Hochheiser and Shneiderman propose *time boxes*, rectangular query regions for specifying queries on time series plots [32]. They act as filters, i.e. time series that exhibit values in the time box' y-range during the time period of interest are selected. Time boxes can also be combined using the Boolean AND operator to formulate more specific queries. Rectangular selection is the brushing mode of our choice when it comes to selecting a subset of curves from a Synchronization Plot. Nevertheless, we realized a limitation of the rectangular brush, in particular in areas where curves are plotted densely: the restriction to an axis-aligned orientation and the rectangle's bounding-box-like characteristics make it highly cumbersome to select exactly the desired set of curves in some situations (Figure 24a and 24b).

This limitation could be overcome by a line brush [45]. Due to its arbitrary orientation and smaller extent, it is more flexible for brushing compared to a selection rectangle. Consequently, it allows to precisely select the desired set of curves, even in dense areas, where a selection rectangle might fail (Figure 24c). A line brush can also simplify the perception of curve progression, where a distinction of the courses of individual curves is challenging due to visual clutter. Crossing a curve by a line brush results in the entire curve being highlighted and consequently catching the analyst's eye. Still, the line brush also has a limitation: due to its nature, curves can only be selected as a whole. Compared to the selection rectangle, there is no possibility to select individual curve sections, which might be helpful for some tasks.

As both brushing mechanisms complement each other in a beneficial way, we implement both modes in our system. During analysis, users can flexibly choose the mode that is most suitable for their analysis task.

#### INTERPRETATION OF BRUSHES

The atomic data unit, on which all views and interactions rely at the most basic level, are time points. A time point corresponds to a data table row, which contains all values measured at that time point. To appropriately reflect a brush in other visualizations, the drawn line or rectangle needs to be transformed into the atomic data unit. However, there is no bijective association between a graphical element, e.g. a curve, and a time point; unlike a scatter plot, where a circle corresponds to exactly one data point.

Consequently, specifying the transformation of a brush to the atomic data unit can be viewed as determining a set of time points that most meaningfully represent the selected (sections of) curves. Remember that the overall purpose of brushing and linking is to relate the brushed curves to patterns and distributions in other views. Independent of the chosen brushing mode, we come up with three interpretations of a brush:

1. *Synchronization Point*: a brushed curve is represented by the synchronization point that is associated with it.

   This approach is supported by both brush modes. A curve might be made up of hundreds of time points. Selecting one of them to represent the entire curve might not be meaningful, when aiming at an exploration of patterns in other views. In this sense, the brushed curve is not adequately represented with this approach.

2. *Entire Curve*: a brushed curve is represented by all time points, of which it is made up.

   This approach can also be used with both brush modes and probably meets the analyst's expectations concerning the representation of a brush in a most intuitive way. The number of selected time points that result from brushing a set of curves might become very large. In this case, it is to be evaluated if corresponding clusters and patterns can be revealed in other views or if they are masked by the sheer number of highlighted data items.

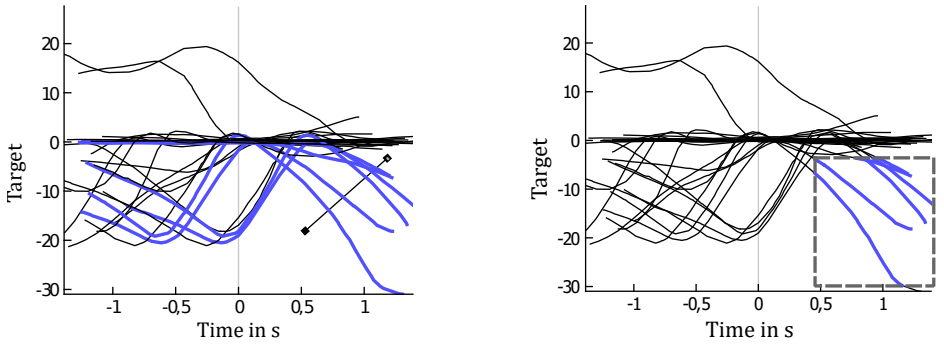
3. *Curve Section*: a brushed curve is represented by those time points that are actually contained in the selection rectangle.

   Obviously, this approach only works for rectangular brushes. It allows the analyst to focus only on that part of the curve that actually incorporates the variance and thus is most interesting to the analyst. No additional, potentially irrelevant time points, which might introduce visual clutter, are considered. Consequently, all further analysis is tailored to the most promising time series parts.

To summarize, the Synchronization Point approach is not suitable for our purpose. The Curve Section approach seems most promising to us, due to its possibility of focusing on exactly those parts of the curves that incorporate the target variance. This interpretation can only be realized together with a rectangular brush. As the line brush provides more flexibility and enhances the perception of curve progression, we offer both brush modes in our system. Depending on the brush mode, the time points marked as selected originate from the entire brushed curves (Figure 25a) or from the curve sections contained in the selection rectangle (Figure 25b).

#### REFLECTING A BRUSH IN THE SYNCHRONIZATION GRID CELLS

Brushing a region of interest in the Focus Plot defines a number of time points, on which further analysis is focused. For each brushed curve section, the corresponding time points are stored successively in the data set. As such, the sequence can be viewed in the global target time series for reference (Figure 26, bottom). Before updating the remaining views in the system accordingly, the brush is handled locally, i.e. within the Synchronization Grid. If a brushed sequence – or part of it – also appears as a curve section in the remaining cells of the Synchronization Grid, this curve section is highlighted in the respective cell (Figure 26, top).




**(a)** *Entire Curve* **(b)** *Curve Section*

**Figure 25:** *Depending on the mode, different sets of time points are considered as selected. The line brush selects the entire curve (a). In contrast, the rectangular brush selects the curve sections within the rectangle (b).*

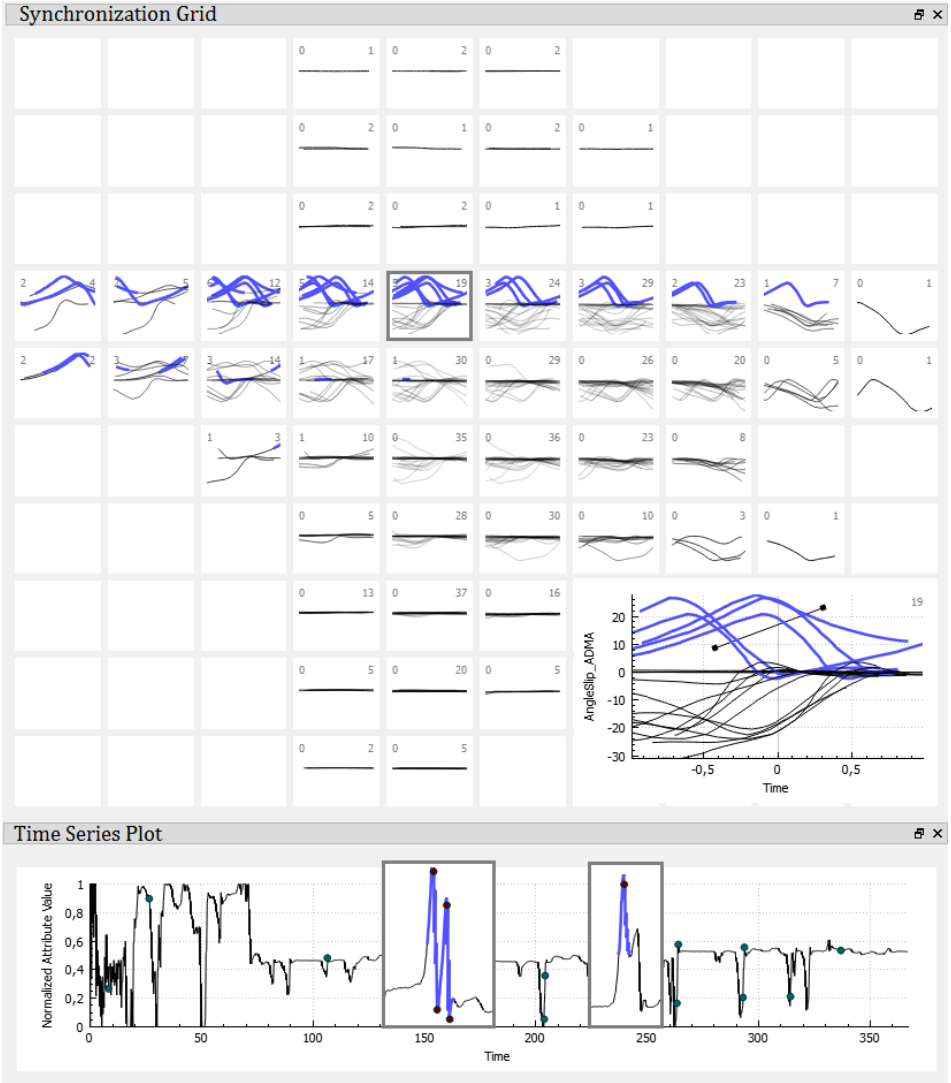



**Figure 26:** *The selected curve sections in the Focus Plot correspond to sections of the global time series (bottom). The brush is handled locally within the Synchronization Grid (top) before updating the remaining views.*

In this way, a linking between the details in the Focus Plot and the cells providing an overview of the feature space is realized. It helps to determine whether the target variance in one cell, which is currently investigated in the Focus Plot, is associated with the variance in other regions of the feature space. When the cells that depict the same brushed curve sections reveal patterns throughout the feature space, this might provide insights with regard to changing initial features. To consider a row (column) of the Synchronization Grid means to observe changes in $X_1(t)$ ($X_2(t)$), while the other feature is fixed to the respective interval determined by the row (column). With this in mind, the internal brushing and linking approach could help analysts to get an impression of whether feature derivations might actually turn out relevant for prediction.

### 5.5.2 *Temporal Context of Synchronization Results*

The curves displayed in each Synchronization Line Plot are actually sections of one continuous target time series. This *parent time series* represents the development of target values over the entire time interval that is covered by the underlying data. Synchronization Line Plots only depict time points in relation to the synchronization points. Thus, the temporal context is not given. The translation of any (relative) point in the Synchronization Line Plot into a point on the parent time series is challenging. Consequently, it requires a large cognitive effort for the analyst to build up a mental image of how all curves of a Synchronization Line Plot are actually part of the same time series. This is, however, important to relate regions of interests and findings to the overall feature and time space.

To support analysts in perceiving this relation, we provide a common *Time Series Plot* to display the parent time series, which is linked to the Focus Plot of the Synchronization Grid. The linking is realized with respect to (1) the synchronization points and (2) the actual curve sections. For the first part, so-called tracers are placed on the parent time series and indicate the positions of the synchronization points as taken from the Focus Plot (Figure 27, circles). However, to fully establish a link between a Synchronization Line Plot and the Time Series Plot, the analyst must also be able to make a connection between curve sections from both plots. Brushing parts of the synchronized curves in the Focus Plot results in the corresponding curve sections being highlighted in the Time Series Plot (Figure 27, lens). This supports analysts in efficiently locating curve sections from the Focus Plot within the parent time series.

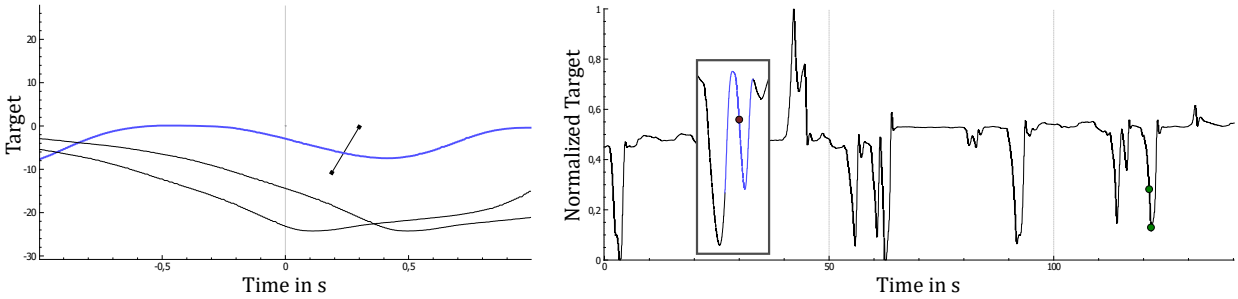



**Figure 27:** *Locating a curve that is brushed in the Focus Plot (left) within the global target time series that covers the entire time range (right). Tracers indicate the positions of synchronization points (circles).*

This connection between synchronized curve sections and the parent time series contributes to an analysis in different ways. Using the Time Series Plot, analysts can get a feeling for patterns in the target behavior that re-occur throughout the global time range. In particular, they can assess how the synchronized curves from a Synchronization Line Plot get in line with such patterns. Secondly, the distribution of a plot's synchronization points across the global time range can be assessed. Analysts can see whether they are equidistant, located close to each other, and whether they cover the entire time range or only certain parts of it. Using brushing, the analyst can also compare the synchronization interval length to the overall time range and to the distance of synchronization points. Being able to evaluate those influence factors highly increases the interpretability of a Synchronization Line Plot, because they raise awareness for characteristics that might otherwise be misinterpreted. Furthermore, the analyst is enabled to notice when the interval length or the synchronization points were chosen badly by the system.

Finally, the Time Series Plot alone also supports an investigation of multiple features at the same time. They are plotted as multiple time series with a common baseline. When used together with highlighting the current synchronization interval via brushing, this functionality is highly useful for evaluating the naïve assumption that the configuration does not change during the synchronization interval. By simultaneously plotting the involved features, the development of their values within the synchronization interval can be observed, allowing for an evaluation of the degree, to which the assumption holds.

### 5.5.3 *Searching for Additional Features*

Our core assumption is that a valid model, i.e. which is based on a valid set of predictors, outputs the same prediction when being given the same input. This characteristic is associated with a low variance of the curves in a Synchronization Line Plot. If the Synchronization Grid contains cells, which contradict this characteristic, this indicates that the underlying feature subset might not be sufficient for explanation of the target feature.

When such a contradicting cell is identified, the analyst wants to find the cause of the varying curves in the corresponding Synchronization Line Plot. One such cause might be another feature being missing in the current model, but which influences the values of the target feature in addition to the considered predictors. In Section 5.5.1, we already introduced the region of interest in a Synchronization Line Plot as the subset of curves that represent the majority of the target variance. Investigating how this region relates to other features can provide valuable hints about the suitability of another independent feature as potential predictor.

#### RELATION BETWEEN VARIANCE AND REMAINING FEATURES

Let us investigate the relation between the region of interest and an independent feature $X_3$, which might be considered as an additional predictor. We refer to the value range of $X_3$ as its domain. We compare the region of interest to a context, which is defined as the entirety of curves within the plot. In the following, a data item refers to a time point. We observe two distributions with regard to $X_3$: (1) the distribution resulting from data items representing the region of interest and (2) the overall distribution, i.e. the distribution corresponding to all data items contained in the plot.

Let us assume the selected cell contains 200 data items in total, of which half of them are brushed as the region of interest. Under the *assumption of independence* between region of interest and independent feature $X_3$, the distribution of the brushed items should follow the overall distribution. This is given when the number of brushed items corresponds to the expected number of items for each part of the discretized domain, e.g. bin of a histogram. In general, the expected number $E(i)$ for a bin $i$ can be determined as $E(i) = \frac{\#_{brushed}}{\#_{total}} * \#_i$ with $\#_i$ as the total number of items in the bin and $\#_{brushed}$ as the number of brushed items and $\#_{total}$ as the total number of items. In our simple case, the expected number for each part is exactly half of the total number of items falling in this bin.

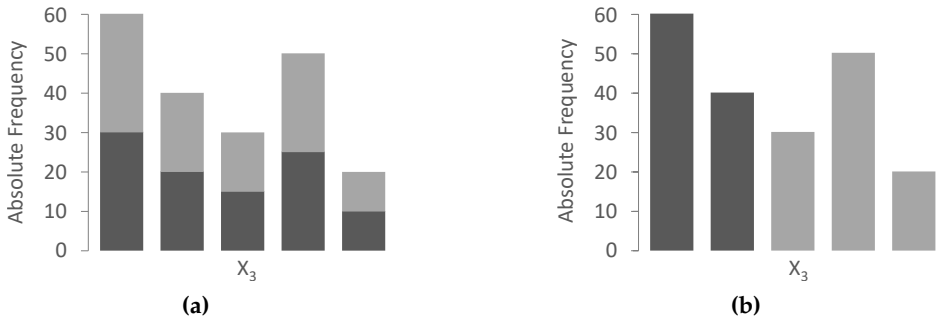



**Figure 28:** *Comparing a region of interest to all curves to assess the relation between predictor candidate $X_3$ and target variance. Independence is given when both distributions are similar (a). A relation exists when the brush distribution deviates from the context distribution (b).*

If the distribution associated with the region of interest follows that of the context (Figure 28a), the hypothesis of independence is confirmed. No particular relation between the region of interest representing the target variance and the independent feature $X_3$ exists. In this case, the independent feature is not assumed to provide valuable information about the target feature and thus is excluded from the set of predictor candidates.

However, if the distribution of brushed items highly differs from the expected distribution (Figure 28b), we reject the hypothesis of independence between target variance and feature $X_3$. If the distribution corresponding to the region of interest even corresponds to a small, coherent part of the domain of $X_3$, this is another strong hint for the target variance being dependent on $X_3$. Including the independent feature $X_3$ into the model might lead to a better discrimination of data items and consequently to a desired less target variance in the newly generated feature space.

#### HISTOGRAM

To support analysts in investigating how a region of interest manifests in different independent features, we link the Focus Plot to a histogram. As we already saw above, it offers a highly intuitive way of assessing the distribution of values over a feature domain. How can a standard histogram be used to view how the target variance manifests in other features? We answer this question with respect to one independent feature to be investigated. However, an exploratory investigation of multiple features can be performed by iteratively checking individual features.

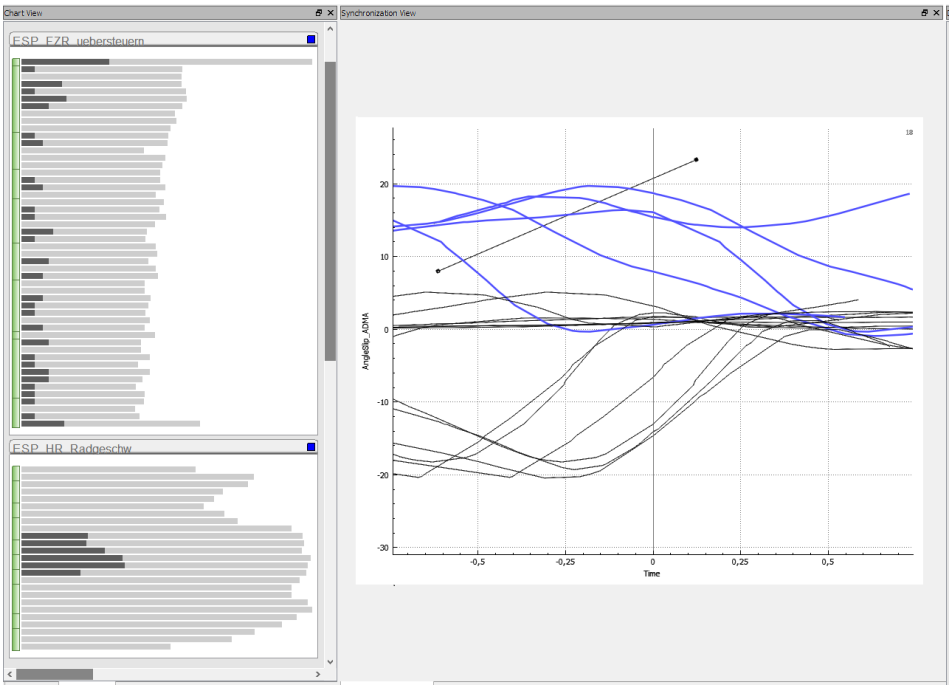



**Figure 29:** *Checking the hypothesis of independence between target variance and two independent features. Histograms (left) are used to compare the brushed curves (right) to all curves as context w.r.t. the distribution of values. The independence hypothesis can be confirmed for the upper feature, while it should be rejected for the lower one.*

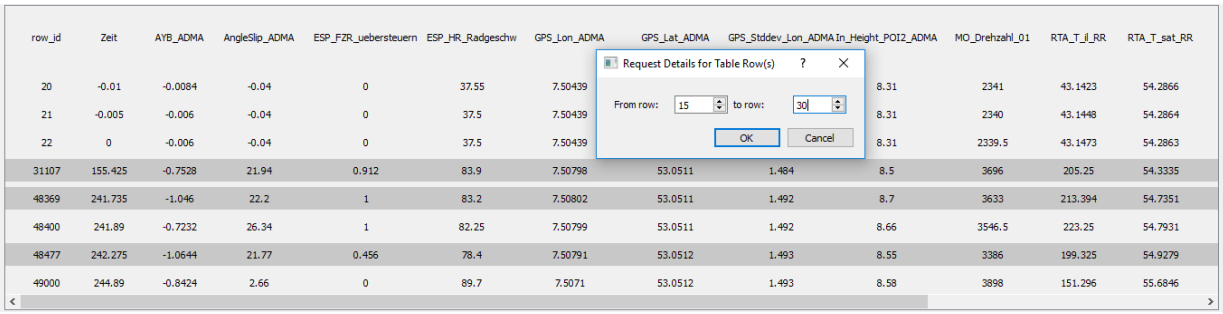
Brushing a subset of curves in a Synchronization Line Plot results in the corresponding time points being marked as selected. Each of these selected time points is associated with a certain value for the investigated feature. Based on this value distribution and the binning of the context histogram, we compute frequencies. These frequencies are then visualized within the context histogram by highlighting a segment of each bar. The segments' lengths are proportional to the number of brushed time points that fall into the corresponding bins. Large time-oriented data sets usually contain ten thousands of time points. However, the number of brushed time points per bin is small compared to the total number of time points contained in that bin. Applying a linear scale to the lengths of segments results in those representing the current brush being hardly visible, because they have a width of only a few pixels. For this reason, we implement a logarithmic scale, which enhances the perception of the small brush segments.

Remember that the brushed curves represent the majority of target variance in the Synchronization Line Plot. By observing their value distribution, the analyst can verify the independence hypothesis by determining (1) which parts of the investigated feature's domain are actually related to the variance and (2) to what extent the brush distribution differs from the context distribution. Figure 29 depicts a brush (right) with its equivalents in the histograms of two different features (left). The differences in the brush manifestations for both features are clearly visible. For the upper feature, no clear mapping between the values corresponding to the brush (dark gray) and parts of the domain can be established. Furthermore, the distribution of the brush in large parts follows the distribution of the context (light gray). This conveys the impression that the feature does not contain useful information about the target. In contrast to that, for the lower feature, the brushed data items only exhibit values in a small range. The variance can be related to this particular part of the feature domain. Also, the distribution of brushed items clearly differs from the overall distribution. This leads to the conclusion that the variance can be reduced by including the feature in the feature subset.

### 5.5.4 *Details on Demand*

Visual representations originate from a mapping of the underlying data. Thus, the analyst perceives only a transformation, which depends on the chosen mapping approach and does not present the data in their original form. It therefore denies access to the raw data. To round off the analysis and to verify findings, the analyst might request details on the exact values that a feature takes on in a specific scenario. For example, after having identified an influencing feature that relates to the variance in a Synchronization Line Plot, one might want to check the feature's exact values for the subset of curves representing this variance. A mechanism is needed, which enables analysts to derive raw data as intermediate results.

The Time Series Plot provides details on demand for displayed features. As a result of its linking to the Focus Plot of the Synchronization Grid, it might contain markers on the displayed time series, which indicate the Focus Plot's synchronization points. As it might be difficult for analysts to read the exact time point associated with a marker from the x-axis, a tooltip is provided, which displays a marker's exact time value when it is hovered. However, the user interaction is limited to requesting exact values for synchronization time points and depicted features.




| row_id | Zeit | AYB_ADMA | AngleSlip_ADMA | ESP_FZR_uebersteuern | ESP_HR_Radgeschw | GPS_Lon_ADMA | GPS_Lat_ADMA | GPS_Stddev_Lon_ADMA | In_Height_POI2_ADMA | MO_Drehzahl_01 | RTA_T_il_RR | RTA_T_sat_RR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | -0.01 | -0.0084 | -0.04 | 0 | 37.55 | 7.50439 | | | 8.31 | 2341 | 43.1423 | 54.2866 |
| 21 | -0.005 | -0.006 | -0.04 | 0 | 37.5 | 7.50439 | | | 8.31 | 2340 | 43.1448 | 54.2864 |
| 22 | 0 | -0.006 | -0.04 | 0 | 37.5 | 7.50439 | | | 8.31 | 2339.5 | 43.1473 | 54.2863 |
| 31107 | 155.425 | -0.7528 | 21.94 | 0.912 | 83.9 | 7.50798 | 53.0511 | 1.484 | 8.5 | 3696 | 205.25 | 54.3335 |
| 48369 | 241.735 | -1.046 | 22.2 | 1 | 83.2 | 7.50802 | 53.0511 | 1.492 | 8.7 | 3633 | 213.394 | 54.7351 |
| 48400 | 241.89 | -0.7232 | 26.34 | 1 | 82.25 | 7.50799 | 53.0511 | 1.492 | 8.66 | 3546.5 | 223.25 | 54.7931 |
| 48477 | 242.275 | -1.0644 | 21.77 | 0.456 | 78.4 | 7.50791 | 53.0512 | 1.493 | 8.55 | 3386 | 199.325 | 54.9279 |
| 49000 | 244.89 | -0.8424 | 2.66 | 0 | 89.7 | 7.5071 | 53.0512 | 1.493 | 8.58 | 3898 | 151.296 | 55.6846 |

**Figure 30:** *Data Table View: providing original records for selected rows. Details are requested via interface or by a brush in another view (dark gray).*

More information can be read from the *Data Table View*, which displays full data table records for selected time points (Figure 30). Time points can be selected either (1) by using the interface, where the analyst can specify a range of rows to be displayed, or (2) by brushing a set of time points in another view (e.g. curve sections in the Synchronization Grid). Requested items are appended to the list of displayed records. If a requested record is already contained, it is highlighted to draw the analyst's attention to it. If needed for a better overview, records can be sorted, resulting in successive time points being displayed below each other. By being able to record data items that they consider worth analyzing, analysts are supported in keeping track of regions of interest throughout the analysis.

## 5.6 IMPLEMENTATION AND SYSTEM INTEGRATION

The proposed Interactive Visual Analysis approach has been implemented as part of a Coordinated Multiple Views system called TableVis. It provides both automated data analysis methods as well as interactive visualizations to support analysis tasks in the context of feature selection. The system is made up of a number of modules, each of which is responsible for one specific task related to handling or visualizing data. The visualization modules include standard multivariate techniques like histogram or scatter plot, but also advanced views that are specifically tailored to the demands of feature selection. All modules are gathered together in the so-called context, which takes charge of scheduling and passing on relevant information between them. Information between modules are communicated in the form of attributes, which store a set of data values and are mostly associated with features of the underlying data. An attribute can be taken as input and modified by a module. All other modules, which take the same attribute as input, are then notified about the changes and

update accordingly. The data, on which the modules operate, are made available in the form of samples. If a module only processes a subset of the available data records, a sample is generated by randomly sampling the underlying data. This might lead to problems when linking two views, because it might not be guaranteed that an item, which is brushed in one view, also exists in another view. For this reason, we introduce a global sample, such that all views operate on the same data base.

The system was implemented using C++ together with the application framework Qt[1] for developing the graphical user interface. The modules are initially positioned inside a main window, but can be re-arranged, resized, docked, and un-docked at any time. The user interface is thus entirely customizable and can be flexibly adjusted to the user's needs. Graphical items are rendered using Qt's built-in Graphics View Framework[2] or its QPainter[3] API. Multi-threading accelerates the processing of expensive operations and contributes to the application's responsiveness.

We use the existing software architecture as described above as a starting point for the development of the techniques presented in this thesis. With the exception of the histogram and scatter plot, the visualizations described in the previous sections were implemented as new modules. The available standard views for multivariate analysis were extended by the Time Series Plot and the Data Table View. The histogram was adjusted to our needs. The core visualization, the Synchronization Grid, consists of several classes that handle the different components. The line plots that are used in the Time Series Plot and in the Synchronization Grid were implemented using a free Qt C++ plotting widget called QCustomPlot[4]. This library was chosen, because it offers a wide range of data visualization and interaction functionalities. The appearance of axes, graph lines, grid lines, and other graphical items can be customized with little effort and functions for processing the manipulation of ranges and plottables are also contained. In particular, the provided selection mechanisms that allow the user to interact with the visual representations were of great help. When it comes to processing user interaction in general, we make extensive use of Qt's Signals and Slots[5] concept. Brushing is realized using a binary attribute to store the resulting selection. This selection attribute can be treated like any other attribute in the system.

1 Qt 4.8: www.qt.io
2 Graphics View Framework: https://doc.qt.io/qt-5/graphicsview.html
3 QPainter Class: https://doc.qt.io/qt-5/qpainter.html
4 QCustomPlot: www.qcustomplot.com
5 Qt Signals and Slots: https://doc.qt.io/qt-5/signalsandslots.html

# 6 CASE STUDY USING REAL-WORLD SENSOR DATA

## 6.1 MOTIVATION

Vehicle dynamics deal with the motion of a road vehicle under the influence of torques and forces. The development of motor vehicles including driving safety, comfort, and assistance systems highly benefits from vehicle dynamics simulation [69]. Because it enables a deeper understanding of the physics of driving, simulation is extensively used to predict the behavior of new vehicles. By enabling an efficient analysis of maneuvers under varying conditions, simulations also reduce tests with real prototypes, allowing for shorter product cycles and lower development costs.

Numerical simulation is based on mathematical models that describe a vehicle's driving behavior. This behavior is influenced by various factors, e.g. the wheel suspensions affect the vehicle stability while changing lanes. Therefore, multi-body models consisting of force elements and rigid elements are commonly used. Research and industry need models that represent the real-world dynamics as accurately as possible. To be able to integrate the simulated systems in various vehicle categories, models should also be able to represent different vehicle classes [78].

The technical background and dataset in this case study are based on the work by Unterreiner, who addresses the precision of multi-body models in representing the input-/output-behavior of a real-world vehicle [77]. A model involves different degrees of freedom (DOF) describing the vehicle components, e.g. orientation of the body or rotational velocities. The transition between such states is given by kinematic equations, which define how the components move in relation to each other. A simulation takes the accelerator pedal position and the steering wheel angle as input. The environment is integrated via predefined conditions, such as side wind or road characteristics [69]. When the simulation has finished, the output quantities characterizing the simulated vehicle's driving behavior are computed from the final system state.

As described above, a generated model should represent real-world driving behavior. Some of the model parameters are already determined by

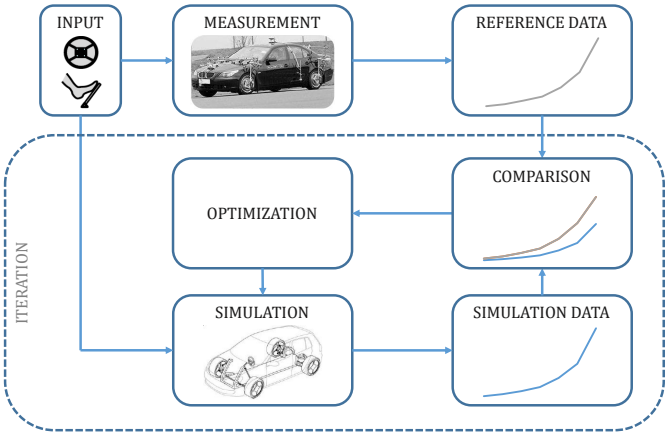



**Figure 31:** *Parameter optimization: the model parameters are iteratively varied such that the simulated vehicle dynamics correspond to the reference behavior from real-world data. Image adjusted from [77].*

specifications for the future vehicle, e.g. the desired engine power. The remaining parameters are determined in an optimization step (Figure 31), where the simulated behavior is adjusted to real-world driving behavior. The reference to reality is given by a reference model, whose driving behavior is recorded by means of various sensors during a test drive.

To yield meaningful results from the parameter optimization, the description of the real-world driving behavior should be as accurate as possible. Missing data within the reference model, i.e. due to a quantity not being properly recorded, poses a problem. During the test drive, the sensor recording the vehicle's *slip angle* over time did not provide a sufficient measurement quality. As a consequence, this quantity cannot be considered for characterization of the real-world driving behavior [77, p. 93].

Experts consider possible solutions: (1) to install a more accurate, optical sensor or (2) to obtain the missing information from knowledge gained from the remaining sensors. An optical sensor is highly expensive and its quality depends on the visibility conditions (e.g. lighting conditions and weather). Thus, instead of installing another sensor, the slip angle values are to be predicted based on existing data. For this purpose, a regression model is needed, which describes the missing quantity as a function of time and of the remaining sensors' temporal development. This involves the task of identifying a minimal subset of sensors, which together are able to sufficiently well predict the missing slip angle measure.

**Figure 32:** *A test car equipped with more than* 100 *sensors.*

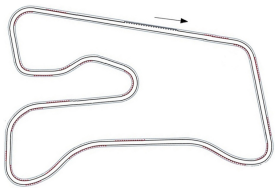


**Figure 33:** *The test track in Germany.*

## 6.2 DATA SET

The sensor data for this case study have been acquired in the context of building a reference model from real-world measurements as described in the previous section. These measurements are registered by means of sensors during a test drive. A car was equipped with more than 100 sensors (Figure 32) and a driver performed a four-lap test drive on the *Handling Track* (HAK) in Papenburg, Germany (Figure 33). The laps were completed in a little more than 6 : 07 minutes and the sensor data were acquired every 5ms, which equals a frequency of 200Hz. We therefore deal with continuous time data, which was discretely sampled. The sensors were already active when the driver passed the leveling area leading towards the actual track. Because these measurements are not representative for the test drive, we exclude the corresponding records from the data set. The resulting data set contains 73,426 items with time stamps covering the range from 0s to 367.125s. Measured quantities include car position, slip angle, rotational speed of the motor, car velocity, yaw rate, and many others. One data item holds the values for all sensors at a specific time (Figure 34).

Detail View

| Time | [illegible] | Slip Angle | ESC_Oversteer | Wheel Speed | GPS_Longit | GPS Latit | [illegible] | Rotation Speed | [illegible] | [illegible] | Yaw Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.105 | -0.0152 | 0 | 0 | 37.6 | 7.50438 | 53.0511 | 8.31 | 2354 | 43.1136 | 54.289 | 0.55 |
| -0.1 | -0.0152 | 0 | 0 | 37.6 | 7.50438 | 53.0511 | 8.31 | 2353 | 43.1142 | 54.2888 | 0.5 |
| -0.095 | -0.0152 | 0 | 0 | 37.6 | 7.50438 | 53.0511 | 8.31 | 2352 | 43.1148 | 54.2887 | 0.45 |
| -0.09 | -0.0004 | 0 | 0 | 37.65 | 7.50439 | 53.0511 | 8.31 | 2350.5 | 43.1154 | 54.2886 | 0.465 |
| -0.085 | -0.0184 | 0 | 0 | 37.7 | 7.50439 | 53.0511 | 8.31 | 2349 | 43.116 | 54.2885 | 0.48 |
| -0.08 | -0.0184 | 0 | 0 | 37.7 | 7.50439 | 53.0511 | 8.31 | 2349 | 43.1166 | 54.2883 | 0.545 |
| -0.075 | -0.002 | 0 | 0 | 37.7 | 7.50439 | 53.0511 | 8.31 | 2349 | 43.1172 | 54.2882 | 0.61 |
| -0.07 | -0.002 | 0 | 0 | 37.65 | 7.50439 | 53.0511 | 8.31 | 2347.5 | 43.1178 | 54.2881 | 0.61 |
| -0.065 | -0.002 | 0 | 0 | 37.6 | 7.50439 | 53.0511 | 8.31 | 2346 | 43.1184 | 54.288 | 0.61 |
| -0.06 | -0.0024 | 0 | 0 | 37.6 | 7.50439 | 53.0511 | 8.31 | 2346 | 43.1191 | 54.2878 | 0.62 |

**Figure 34:** *10 items of the data table depicting the sensor data. One item is associated with one time point and holds values for all sensors.*

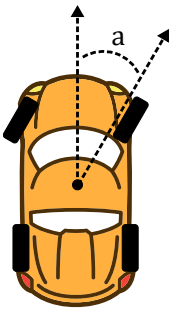



**Figure 35:** *Slip angle between directions towards which body and tires are pointing.*

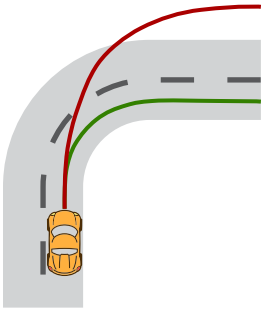


**Figure 36:** *Understeering: the vehicle steers less than the amount intended.*

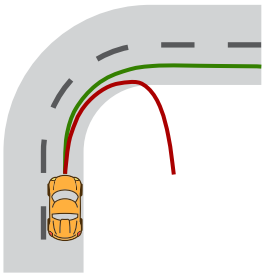


**Figure 37:** *Oversteering: the vehicle turns more sharply than commanded.*

## 6.3 FEATURE SELECTION FOR TIME-DEPENDENT SENSOR DATA

In the previous section, we already explained that the slip angle sensor does not provide the required measurement quality. Its values should therefore be predicted based on the existing sensor data. As a consequence, slip angle will be the target feature throughout our analysis. By exploring the described multivariate, time-dependent data set, we then search for a minimal descriptive subset of the remaining sensors, on which regression modeling is promising for prediction.

### 6.3.1 *Target Feature: Slip Angle*

In vehicle dynamics, the slip angle refers to the angle between the direction that the vehicle is actually traveling (i.e. towards which the tires are pointing) and the direction that the vehicle body is pointing [63]. This situation is depicted in Figure 35. A non-zero slip angle mainly occurs during cornering. It results in a force perpendicular to the tire's direction of travel. When this force is larger than the tire's friction resistance, the tire will start to move sideways. The ratio between the front and rear slip angles determines the driving behavior of a vehicle during a curve. If the slip in the front wheel is greater than the rear slip, the vehicle will *understeer* [17]. As a consequence, the vehicle will steer less than the amount intended by the driver and potentially leave the road (Figure 36). If the rear slip is greater than the slip in the front wheel, the behavior will be characterized by *oversteering*. This results in the vehicle turning more sharply than commanded (Figure 37).

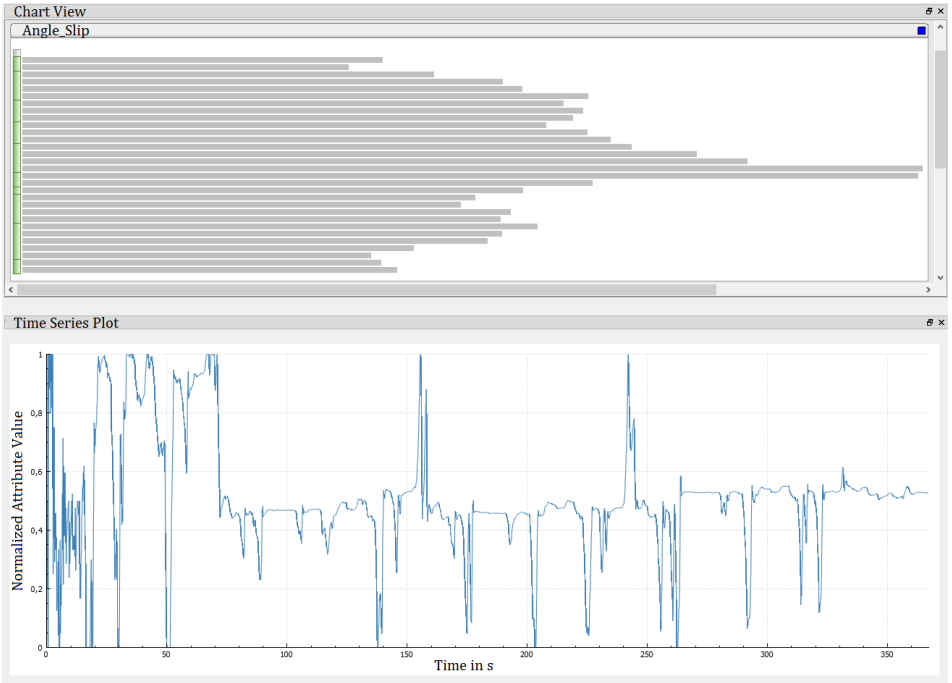



**Figure 38:** *Starting point: gaining an overview of the target feature. A histogram depicts the shape of its value distribution (top). Temporal context is added by depicting the value development as a time series (bottom).*

### OVERVIEW OF ITS DISTRIBUTION

Before dealing with particular analysis tasks and performing a target-oriented exploration of the time-dependent features, the analysis starts by gaining an overview of the target feature. First of all, we are interested in the overall distribution of its values. At this stage, we do not care about the points in time, at which the values arose. Instead, we simply want to get an impression of which values the target feature takes on at all.

A well-known visualization that allows for an intuitive perception of the main characteristics of a distribution is the box plot. It summarizes large data sets using five measures: the minimum, lower quartile, median, upper quartile, and maximum. As such, it is particularly suited for comparing multiple distributions. However, at this stage, we are particularly interested in the distribution of one single feature – the target. In this context, the five-number summary of a box plot provides an intuitive first impression, but rather little detail. In contrast to that, a histogram better addresses the visual sense by approximating the actual shape of the distribution. The slip angle is measured relative to the direction of the vehicle

body. Negative values denote that the wheels are steered to the right and vice versa. In Figure 38, top, we can see that the distribution is unsurprisingly symmetric and not skewed. It is approximately bell-shaped, but still reveals five local peaks: one on either end of the value range, the global maximum in the center, and one on either side of the maximum. Due to the slip angle's symmetric value range, the global maximum frequency arising at a value of approximately zero seems legit. It can be related to straight sections of the test track, of which there are more than curved sections, which are related to a non-zero slip angle. However, the histogram only depicts aggregated frequencies and does not convey statistical values like minimum and maximum of the target distribution.

#### OVERVIEW OF ITS TEMPORAL DEVELOPMENT

Until now, we investigated the target feature from the perspective of its values, independent of when those values arose in time. However, a thorough overview also includes the temporal perspective. For this purpose, we employ the well-known Time Series Plot (Figure 38, bottom). The attribute values are normalized to $[0, 1]$ to retain the possibility of depicting and comparing multiple time series. As we characterized the target distribution as a rather symmetric one centered around a slip angle of zero from the histogram and Dash Plot, we can now mentally map the zero slip angle to a normalized value of 0.5. The Time Series Plot shows rather stable time series sections around $y = 0.5$ in between sharp peaks. The former can be related to straight sections in the test track, while the latter might correspond to curves. The four laps that have been performed on the test track can be clearly identified as a re-occurring pattern. During the initial phase (i.e. the first minute), this pattern is rather unsteady, before it becomes stable from the second lap onwards.

### 6.3.2 *Initial Features: Yaw Rate and Vehicle Velocity*

The overall analysis task is to find features that contain valuable information about the slip angle, which make them suitable predictors. Besides analyzing individual features independently, we therefore have to investigate the relationships between the target and the remaining features. The slip angle is the difference between the heading direction of the vehicle body and the direction, in which the tires are pointing. It is therefore reasonable to assume that the slip angle is related to the yaw rotation, i.e. the change of the direction towards which the vehicle body is pointing. It is captured with a sensor measuring the *yaw rate* of the car. When comparing the time series of both slip angle and yaw rate (Figure 39), we

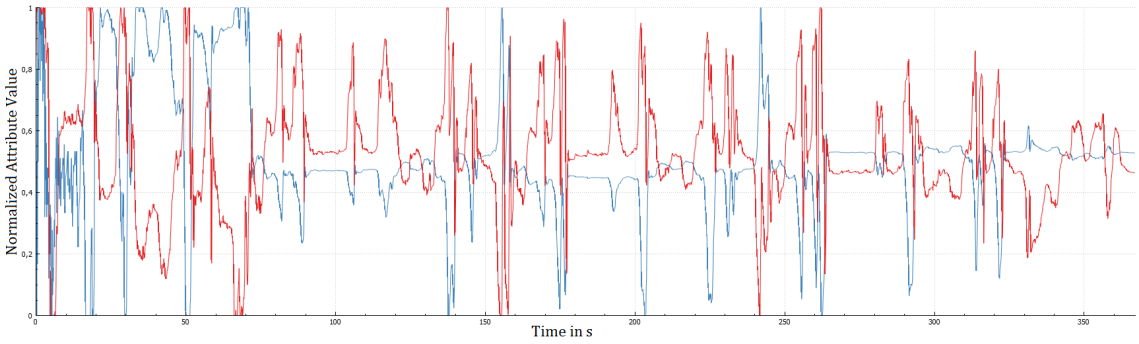



**Figure 39:** *Investigating potential predictors: yaw rate (red) is assumed to be related to slip angle (blue). Simultaneously occurring characteristics in the Time Series Plot strengthen this hypothesis.*

notice that peaks and uniform sections in both time series occur simultaneously, i.e. without latency. The majority of these characteristics in the target time series can thus be associated with a counterpart in the yaw rate time series. Still, although both features seem to be strongly related, simply horizontally mirroring the yaw rate time series does not yield a precise description of the slip angle's temporal development.

As a consequence, we search for an additional feature that might provide information for the slip angle prediction. From our knowledge in the domain of vehicle dynamics, we conclude that the vehicle's *velocity* also influences the slip angle, as speed also contributes to the emergence of forces applying in directions other than the traveling direction. In addition, both the yaw rate and the velocity of a vehicle are features that do not depend on other measures within the investigated system, but are solely determined by an external factor, i.e. the driver turning the steering wheel and specifying the accelerator position. Such features are a good starting point for the identification of a minimal descriptive set of predictors, because they cover characteristics of the target feature that no other sensor can explain. Furthermore, we can assure that their relations to the target are not an artifact of a common cause.

### 6.3.3 *Explanatory Power of the Initial Feature Subset*

As a start, we have gained a first impression of the target feature and used our domain knowledge to determine an initial set of potential predictors. In the following analysis, we aim at an estimation of the feature set's meaningfulness for the prediction of the target feature.

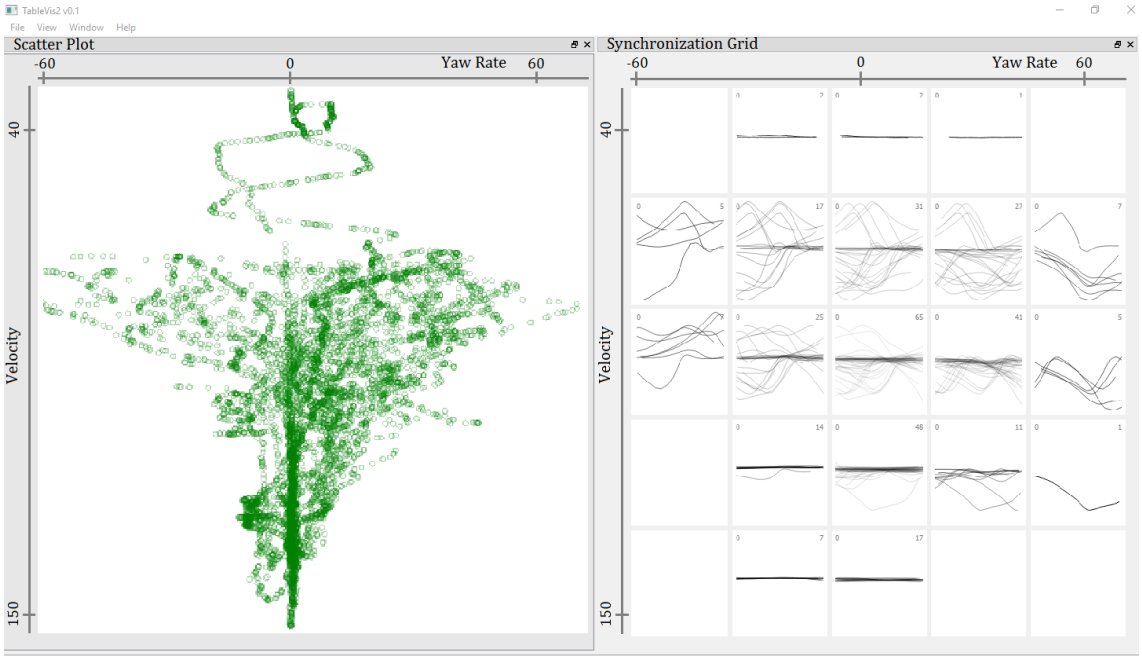



**Figure 40:** *Overview of the initial feature space determined by domain knowlegde. Physically unlikely value combinations are not present (left). The Synchronization Grid conveys a first impression of the feature subset's explanatory power (right).*

### OVERVIEW OF THE FEATURE SPACE

Using a scatter plot, we first take a look at the distribution of data points within the feature space, which is spanned by yaw rate and velocity (Figure 40, left). Note the empty areas, where the corresponding value combinations are not present. For example, the empty regions in the lower corners indicate that high velocity does not occur together with a high absolute value of yaw rate. This can be explained by physical constraints, as a vehicle would leave the road when taking a turn with too high velocity. Accordingly, the highest velocities only show up together with a yaw rate close to zero, indicating straight sections of the test track, which allow for high speed extension. Such conclusions can be intuitively drawn based on physical understanding, but visualizing the feature space as a scatter plot is still helpful for building up a mental model, to which analysts can go back whenever they examine more abstract visualizations.

Such a complex visualization is the Synchronization Grid, which we take into account to get an idea of the explanatory power of the feature subset $\{yawrate, velocity\}$. We start with a grid resolution of five cells in

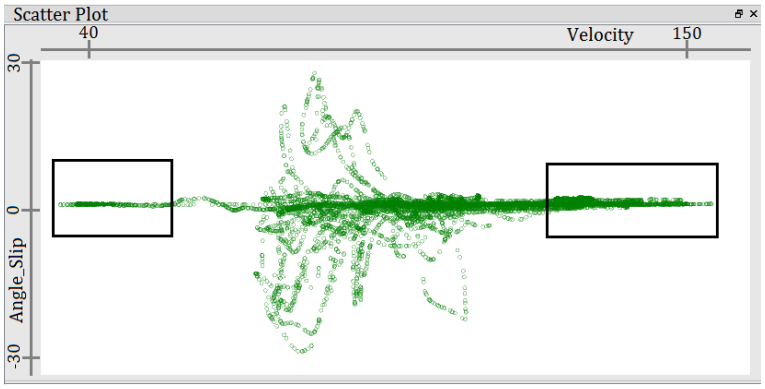



**Figure 41:** *Plotting slip angle against vehicle velocity confirms the hypothesis of velocity being a suitable predictor for specific parts of the data. The scatter plot reveals an almost functional relationship for items with rather low or high velocity (rectangles).*

both directions. Such a low resolution does not provide much detail, as a large number of data points within each cell is aggregated to a representative configuration. Still, it conveys a notion of how the target feature develops around those configurations (Figure 40, right). Remember that a valid model would predict similar target developments based on similar configurations, i.e. those configurations summarized within one cell. The first and last rows of the Synchronization Grid show Synchronization Line Plots, where all curves are nearly equal. From this, we can conclude that the vehicle's velocity is a suitable predictor for the slip angle for these portions of the feature space, i.e. when considering only data records with very low and very high velocity values. Plotting the slip angle against the velocity in a scatter plot confirms this hypothesis, as a relation is clearly visible from the arrangement of the data points. In fact, it reveals a strong linear correlation for a subset of data records (Figure 41).

#### REGION OF INTEREST

The vehicle velocity is only helpful to predict a part of the data. Thus, we re-consider the Synchronization Grid as an overview of the velocity together with the yaw rate for further analysis. As we can locate cells, for which the curves exhibit highly varying shapes, this indicates that the two features are not descriptive enough to cover all characteristics of the target feature. To identify critical regions of interest for a more detailed analysis, we first need to narrow the feature space to areas with an undesired target variance. For this purpose, we double the resolution of the Synchronization Grid (Figure 42). The resulting detailed representation

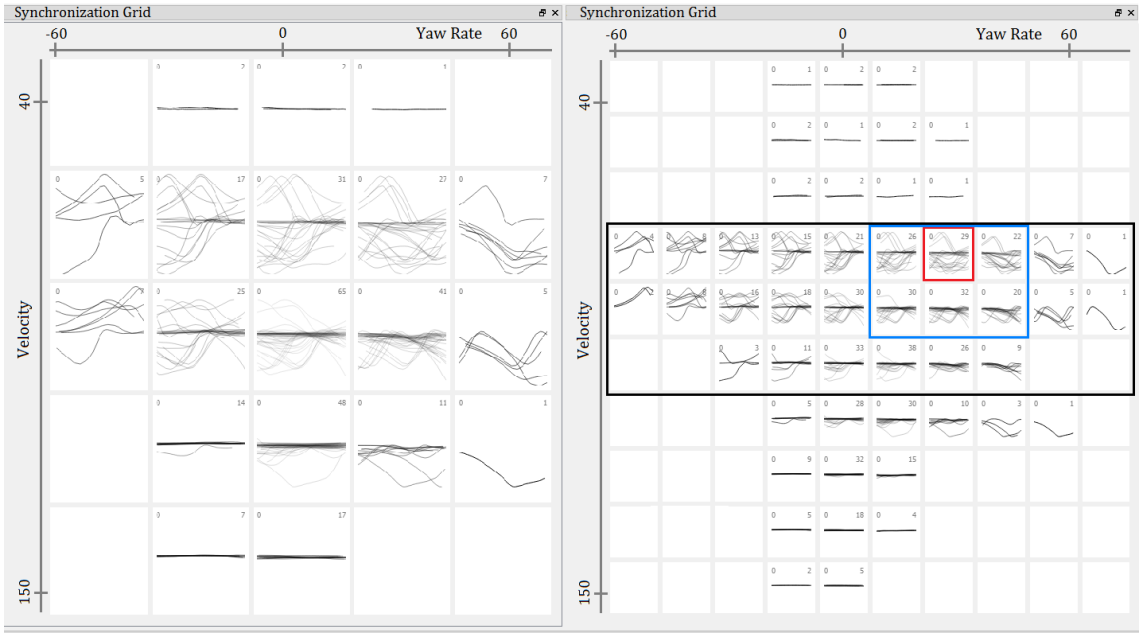



**Figure 42:** *Increasing the grid resolution for details. We focus on cells exhibiting undesired variance (black rectangle). The critical region of interest are those cells, where the curves are perceived as most varying after the synchronization point (i.e. for $x > 0$) (blue rectangle). One cell is selected for further analysis (red rectangle).*

of the feature space reveals that the valid model assumption is not fulfilled for nearly one third of the feature space (Figure 42, black rectangle). Many cells within this region of interest exhibit a significant target variance, although each of them also contains a number of curves behaving similarly. In the cells of the Synchronization Grid overview, these subsets of similar curves are visible as dark horizontal strokes, which result from the semi-transparent curves adding up to black.

As a next step, we intend to identify the cell with the least quality, i.e. the largest target variance. An in-depth analysis of this cell is the most straight-forward way to clarify the circumstances, under which such high target variance arises. It can also be the starting point for searching the actual cause of the variance, which might be an additional feature, on which the target depends. However, the target variance arises in different forms and the decision for the one cell with the highest variance might be an unclear task. In the Synchronization Grid, we can distinguish between cells, where (1) the variance occurs everywhere, but its spread is

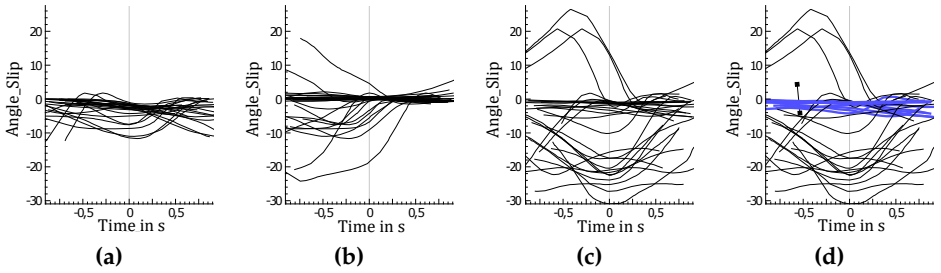
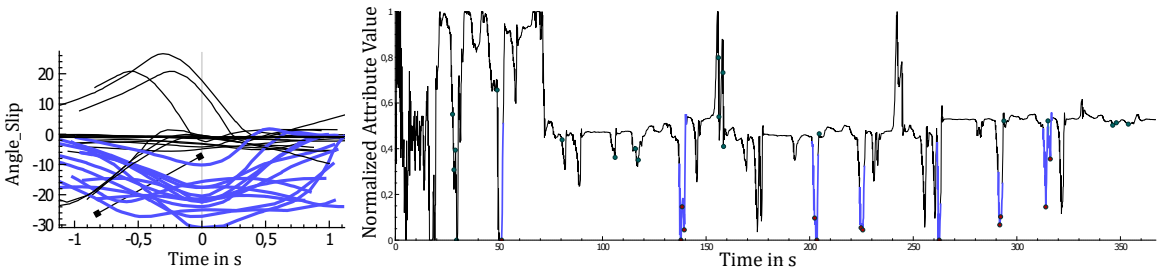



**Figure 43:** *Different forms of variance: curves only spread within a small* $y$*-range (a), differing behavior becomes consistent after the synchronization point (b), or each curve takes a different course (c). Even cells with large variance contain a set of similar curves (d).*

limited (Figure 43a), (2) the variance occurs mainly before the synchronization point (i.e. where $x < 0$) and then evens out (Figure 43b), and (3) the curves occupying large parts of the target range (Figure 43c). Cells where each curve takes a highly different course can be said to have a lower quality than those with an acceptable spread or where curves develop towards a consistent behavior. With this ranking in mind, we can narrow the critical region of the feature space to those cells marked with blue rectangles in Figure 42. The presence of such a critical region indicates that the overall explanatory power of the underlying feature set is not sufficient for prediction. Consequently, we select one of the most critical cells for a more detailed analysis (Figure 42, red rectangle). Detailed analysis is performed by exploring and interacting with the corresponding Synchronization Line Plot as a Focus Plot. It involves the search for an additional feature that is related to the target variance and thus would increase the explanatory power when added to the feature subset.

### 6.3.4 *Searching for Another Influencing Factor*

Based on the feature subset $\{yawrate, velocity\}$, we aim at identifying the next feature to add, such that the overall explanatory power is increased. For this purpose, we investigate that region of the feature space, for which the prediction is most unstable, i.e. the cell selected in the previous step. Our goal at this stage of the analysis is to explicitly identify a feature, to which the variance in this region relates. Adding this feature to the feature subset might result in an improved model, which explains more characteristics of slip angle as the target than could be described by yaw rate and velocity alone. However, there might exist (1) inter-dependencies




**(a)** *Region of Interest* **(b)** *Temporal Context*

**Figure 44:** *The decision which feature to add next focuses on those relating to the remaining target variance. We start by brushing the curves representing the majority of variance (a). Their locations within the global target time series are depicted for reference (b).*

between the identified third feature and the features already contained or (2) a relationship between the variance and changes in the initial features. In both cases, adding the identified feature to the feature subset might not yield the expected improvement of the explanatory power. For this reason, we have to carefully examine the true impact that adding the feature has on the distribution of the target variance.

#### REGION OF INTEREST

The key question that needs to be answered when deciding for the next feature to be added is the following: how do those curves incorporating the target variance within the cell – i.e. the region of interest according to Section 5.5.1 – manifest in other features?

We first define the region of interest in the corresponding Focus Plot via brushing. In our case, the plot exhibits three curves in the upper half of the plot, a number of curves behaving as desired around $y = 0$, and a number of highly varying curves in the lower half of the plot. Previous analysis of the three upper curves using brushing and linking together with the Time Series Plot revealed that these curves actually originate from replication. Thus, they contain redundant information and are not of interest to us. Consequently, the set of curves occupying the lower half of the Synchronization Line Plot is brushed as the region of interest (Figure 44a). Its placement within the global temporal context is depicted in Figure 44b. To identify a potential influencing factor, we now analyze the remaining features in the dataset with regard to whether they relate to the region of interest.


**Figure 45:** *Evaluating four potential predictors regarding their influence on the target variance. ROI (top) and context (bottom) used for comparison (a). Differing distributions w.r.t. the three leftmost features indicate an influence. The rightmost feature is considered independent, as ROI and context distributions are similar (b).*

HISTOGRAM: WHICH FEATURES RELATE TO THE VARIANCE?

Histograms allow us to use our visual sense to efficiently compare the distribution associated with brushed curves to the distribution of all curves in the cell. A feature is assumed to relate to the target variance represented by the brush, if (1) the distribution of the brush highly differs from that of the context and (2) the brushed data items are distributed only within a self-contained part of the feature's domain. Note that the latter is not a necessary condition. A suitable predictor might be found among features, for which brushing the target variance influences the shape and domain coverage of the distribution.

Figure 45b provides an overview of histograms depicting the region of interest (top) and context (bottom) distributions for four different features (columns) on a logarithmic scale. The curves corresponding to the region of interest and context are given for reference (Figure 45a). For the three leftmost features, the brush distribution clearly differs from the context distribution. For now, we can conclude that a relation between those features and the investigated target variance exists. Including the three features in the current feature subset might add significant value to it.

Still, when investigating the corresponding histograms in more detail, we can perceive different characteristics concerning the deviation of the brush distribution from the context. These characteristics give us subtle hints concerning the strength of the relation between feature and target variance. For the first feature, the difference between both distributions primarily consists in one of the two peaks in the context distribution. However, the brush distribution still exhibits a similar spread. In this sense, the influence of the region of interest on the shape of the distribution is not as significant as for the second feature. There, the brush manifests within a small part of the feature domain, indicating that the variance can be strongly related to this same portion of feature values. In contrast to that, the context distribution covers the entire domain. The brush distribution seems to be made up of the thinner one of the context distribution's two separable parts. Another interesting case can be observed for the third feature. Here, both distributions cover approximately the same part of the feature domain. The difference consists in the gaps that the brush distribution exhibits, which are not present in the context distribution. In this way, both distributions significantly differ, but – unlike for the second feature – the region of interest and therefore the target variance cannot be associated with a separate part of the feature domain. In contrast to the features possibly related to the target variance, the rightmost feature does not show changes in the distribution when considering the two different sets of curves. The two corresponding histograms show that the distribution of the brush mainly follows the overall distribution.

According to our independence criterion, the rightmost feature is not worth to be considered as a predictor candidate. The varying distributions for the remaining three features lead towards the conclusion that all of them might be related to the target variance and thus suitable as an additional predictor to be included in the model. However, the degree and appearance of the differences between brush and context distributions provide an additional impression of the strength of this relation. For the second feature, the brush distribution occupying one part of the domain is complementary to the remaining part of the context distribution. At the same time, the brush only occupies a small range of the feature's domain. Both are strong hints for the feature being related to the variance represented by the brush. We therefore conclude that the second feature has the strongest relation to the target variance and thus should be added to the feature subset to improve the predictive power. Nevertheless, conclusions concerning the existence or strength of a relation need to be handled with care and should be double-checked.

# 7

# PROOF OF CONCEPT USING AN ARTIFICIAL DATA SET

The case study in Chapter 6 showed that the proposed system can successfully be used to perform the first iteration of a feature selection procedure for time-dependent sensor data. As our approach was developed to be non-sensitive to specific characteristics of the underlying data, we believe that it is applicable to any numerical time-dependent data.

To evaluate the methodology's applicability in a general context, we provide a Proof of Concept that is based on an artificial data set. To mimic a real-world setting, where the analyst does not know the data, this informal evaluation involves two entities: (1) the constructor, who generates the time-dependent data with preconceived patterns, and (2) the analyst, who does not have any knowledge about the patterns. The analyst then works out insights into the data, while the ground truth is given by the generating equations that the constructor invented.

The Proof of Concept conveys an impression of the functionality, but also the limitations, of the proposed IVA system. Being able to compare the analysis findings to the ground truth also allows for assessing the level of detail that can be achieved. This could be the degree to which an identified feature subset matches the ground truth. The analysis procedure consists of two steps: (1) a general exploration of the features and patterns without any previous knowledge about the data set and (2) the first step of a feature selection given two initial features and a target feature.

## 7.1 GENERAL EXPLORATION WITHOUT PREVIOUS KNOWLEDGE

The data set contains 18 features, whose values were recorded for $10,000$ equidistant time points. To mask the meaning of the features, they are labeled with the letters a to r. The following exploratory analysis begins with an investigation of the value distributions and temporal developments for each feature separately. At the next stage, multiple time series are plotted simultaneously to search for co-occurring patterns and thus potential relationships. Finally, the explanatory power of a chosen feature subset is analyzed using the synchronization approach.

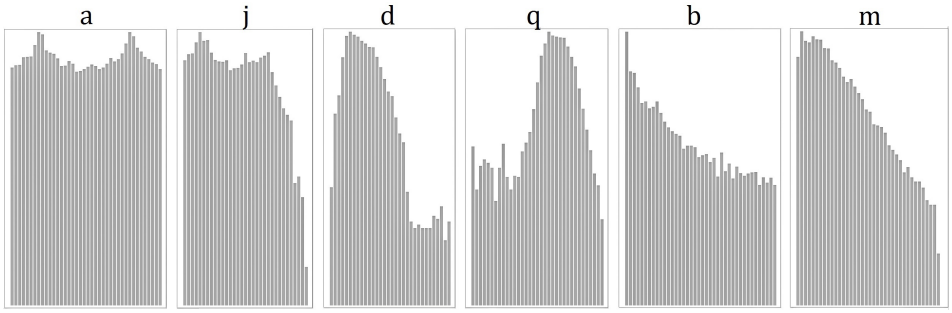


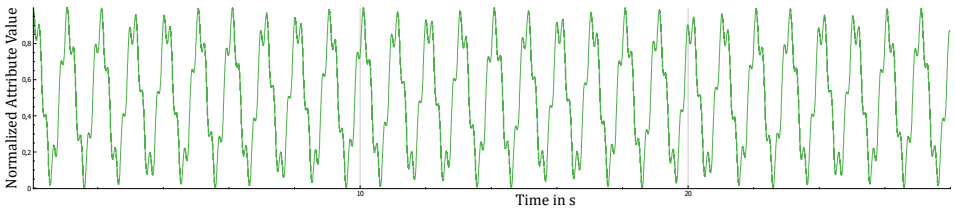

**Figure 46:** *A selection of features with outstanding distribution characteristics as starting points for an exploratory analysis of patterns in the data.*

### 7.1.1 *Value Distribution*

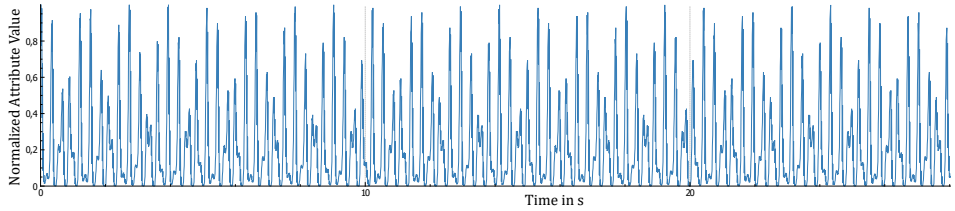
We start the actual analysis by investigating each feature's value distribution. We primarily search for any outstanding characteristics that might pose a starting point for further analysis. For an exploratory analysis, there can be multiple such starting points, each of which might lead towards another insight. Points of interest could be features with a remarkable distribution, e.g. one that might be non-symmetric, bi-modal, or skewed. Figure 46 shows six selected histograms. Besides providing an overview of feature values, histograms can also raise a first awareness for features whose time series potentially relate, as similar or mirrored histograms might be produced by identical (and potentially phase-shifted) or inversely developing time series. We identified feature i to exhibit only one value. Thus, it does not contain valuable information and is excluded from further analysis. We also noticed that some histograms were highly similar. Comparing the corresponding time series helps to determine whether these features indeed behave in a similar way.

### 7.1.2 *Temporal Development*

For such purposes, we consider the Time Series Plot next. It stands out at first sight that the data is highly periodic. Observing the shapes of time series might be helpful for determining features that are not at all influenced by components of the system. Their values solely depend on time as well as the values that the same feature took on at previous time points. Due to this property, the time series of an independent feature might exhibit rather undisturbed and uniform oscillations, which do not exhibit sudden changes that are likely to be caused by an influencing feature. Such characteristics can be observed for features a (Figure 47a)




**(a)** *Feature* a




**(b)** *Feature* b

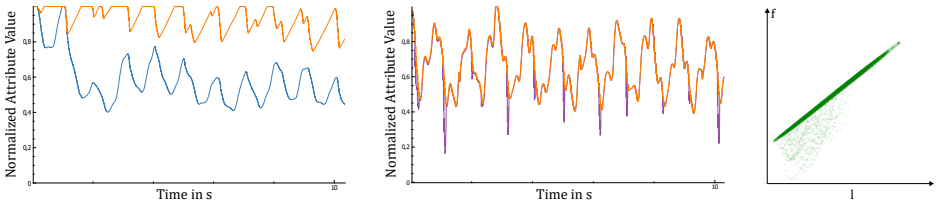
**Figure 47:** *Two features whose undisturbed periodic oscillations suggest that they are entirely independent of the investigated system and thus might be a good starting point for the identification of relationships.*

and b (Figure 47b).Independent features are important, because they do not pose the risk of being an intermediate variable or having a common cause induce a relation to another feature.

**Finding 0.1 (Independent Features)** *Features* a *and* b *might be independent of any other feature in the investigated system.*

In a second step, we compare multiple time series and look out for interrelations. The overall goal at this stage is to build up feature subsets for a more detailed analysis using the Synchronization Grid. One way of identifying such feature subsets is to plot multiple time series simultaneously and investigate their concurrent development. Particular attention is paid to whether peaks or valleys in different time series co-occur or whether one time series increases as another one decreases.

We reconsider the features with remarkable histograms as a starting point. The Time Series Plot shows that the time series of features d and g as well as l and r are actually exactly the same. As a consequence, we exclude g and r from further analysis to avoid the redundancy introduced by those pairs. The time series of features a and j are strongly related in terms of the locations of peaks and valleys as well as the curve slope. For a correct interpretation of future findings, one should keep in mind that j might originate from a in large parts (as we assume a to be independent).




**(a)** d *(blue) and* q *(orange)* **(b)** l *(orange) and* f *(purple)* **(c)** l *and* f

**Figure 48:** *Investigating relationships between features. We assume* d *to depend on* q*, as a decrease in* q *is mostly followed by a decrease in* d *(a). Large parts of* f *can be explained by* l *(b). The scatter plot confirms a strong linear correlation except for a cloud of outliers (c).*

**Finding 0.2 (Feature Equality)** *The data set contains two pairs of equal features:* d = g *and* l = r*. Features* g *and* r *are excluded from further analysis. Feature* j *is assumed to be highly dependent on feature* a*.*

From the histograms of features d and q, we could conclude that their time series might be approximately mirrored. At first glance, this hypothesis can be confirmed in the Time Series Plot (Figure 48a). However, when taking a closer look at the behavior, we can see that – for most peaks of d – the decrease of d-values starts simultaneously to or slightly after a decrease in q. Due to the temporal order of decreases, we might assume that changes in d are influenced by the behavior of q. In Section 7.2, we consider an additional influencing feature as given and analyze the resulting feature subset as a showcase using the Synchronization Grid.

**Finding 0.3 (Influencing Factor)** *Changes in feature* d *are assumed to be influenced by the behavior of feature* q*.*

For features f and l, the Time Series Plot reveals that they are almost equal, except for some sharp valleys of f (Figure 48b). The scatter plot of both features shows a strong linear correlation, with a cloud of outlier points at the lower end, where those valleys manifest (Figure 48c). The question is: which feature(s) can help to explain those valleys in f that l does not cover? Further exploration of f together with different features in the Time Series Plot reveals that the time series of k contains highly similar valleys at exactly the same locations (Figure 49). From this analysis, we conclude that l and k each explain complementary parts of f.

**Finding 0.4 (Two Features Explaining One Target)** *Feature* f *is assumed to depend on* k *and* l*. Both features seem to explain complementary parts of* f*.*

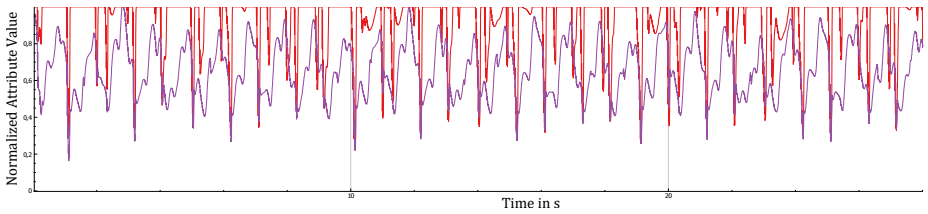
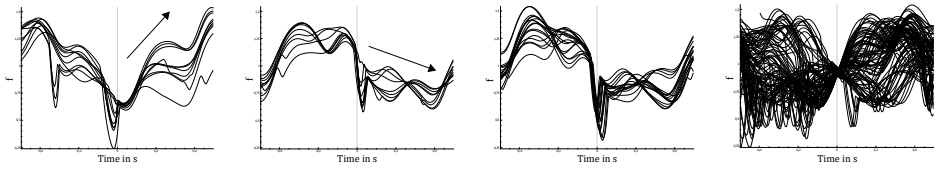



**Figure 49:** *Feature* k *(red) exhibits sharp downwards peaks that approximately correspond to those of* f *(purple), which are not covered by* l *(Figure 48b). Consequently,* k *and* l *explain complementary parts of* f*.*




**(a)** **(b)** **(c)** **(d)**

**Figure 50:** *The effect of varying* l *and* k *on the target curves. Feature* l *influences the slope. Low values are associated with a positive gradient (a). High* l*-values lead to decreasing target values (b). Feature* k *influences the variance. Low values result in little variance (c). Large* k*-values are critical, as they co-occur with strongly differing target behaviors (d).*

### 7.1.3 *Searching for an Additional Feature*

After having performed the exploratory analysis, we now verify the hypothesis of l and k together explaining feature f. For this purpose, we take a look at the synchronized behavior of the target f across the feature space spanned by l and k. We first notice the effects of varying l and k on the target behavior. The slope of the target curves is considerably influenced by the values of l (Figure 50a and 50b). Furthermore, feature k has a notable effect on the degree of variance exhibited in the cells. This variance increases with increasing k (Figure 50c and 50d). We believe that this observation is plausible, as k only explains the large valleys of f, where its own values are rather low (recap Figure 49). The remaining development of f-values is supposed to be described by l.

As the Synchronization Grid exhibits undesired variance for several cells, we take k and l as predictors and search for a third independent feature that might relate to these differences in the synchronized behavior of the target f. We perform a detailed analysis of one of the cells with large vari-

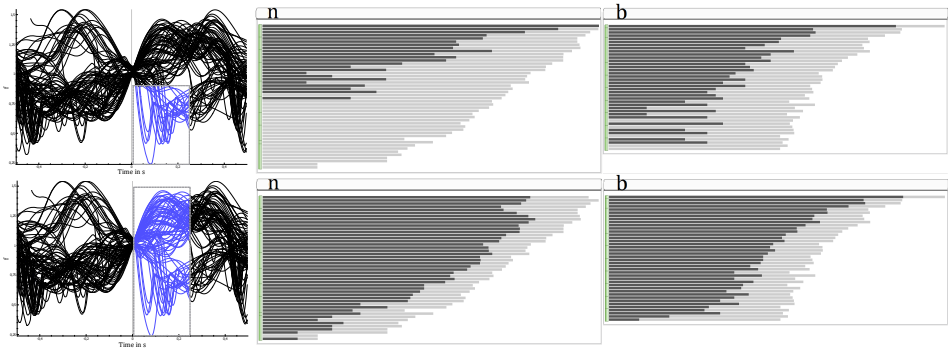



**Figure 51:** *Do* n *or* b *relate to the target variance? Curves representing the variance are compared to all curves as context (left). For* n*, the distribution associated with variance clearly deviates from the context, indicating a relationship (center). Brushing the variance does not have an effect on the distribution of* b*, indicating independence (right).*

ance. It contains two sets of diverging curves. One is brushed as region of interest, while a context brush covers all curves in the same time range (Figure 51, left).

We compare the distributions resulting from those brushes for two features n and b. For feature n, brushing the region of interest clearly influences the distribution's shape (Figure 51, center). The brushed region solely covers the upper half of the feature domain, while the distribution associated with the context brush is spread across the entire domain. In contrast to feature n, the distribution for b is not significantly influenced by brushing the region of interest (Figure 51, right). Both histograms exhibit a similar shape and cover the whole feature domain.

These findings suggest that feature n could be meaningful when it comes to further discriminating between both sets of curves and thus reducing the target variance. Feature b would be less helpful for this purpose. As a conclusion, n is worth considering as an additional predictor candidate. However, its interdependencies with the current feature subset as well as the actual benefit of additionally including it into the model have to be evaluated in more detail.

**Finding 0.5 (Additional Feature)** *Feature* f *cannot be sufficiently explained by* k *and* l*. Feature* n *was determined as a suitable additional predictor candidate.*

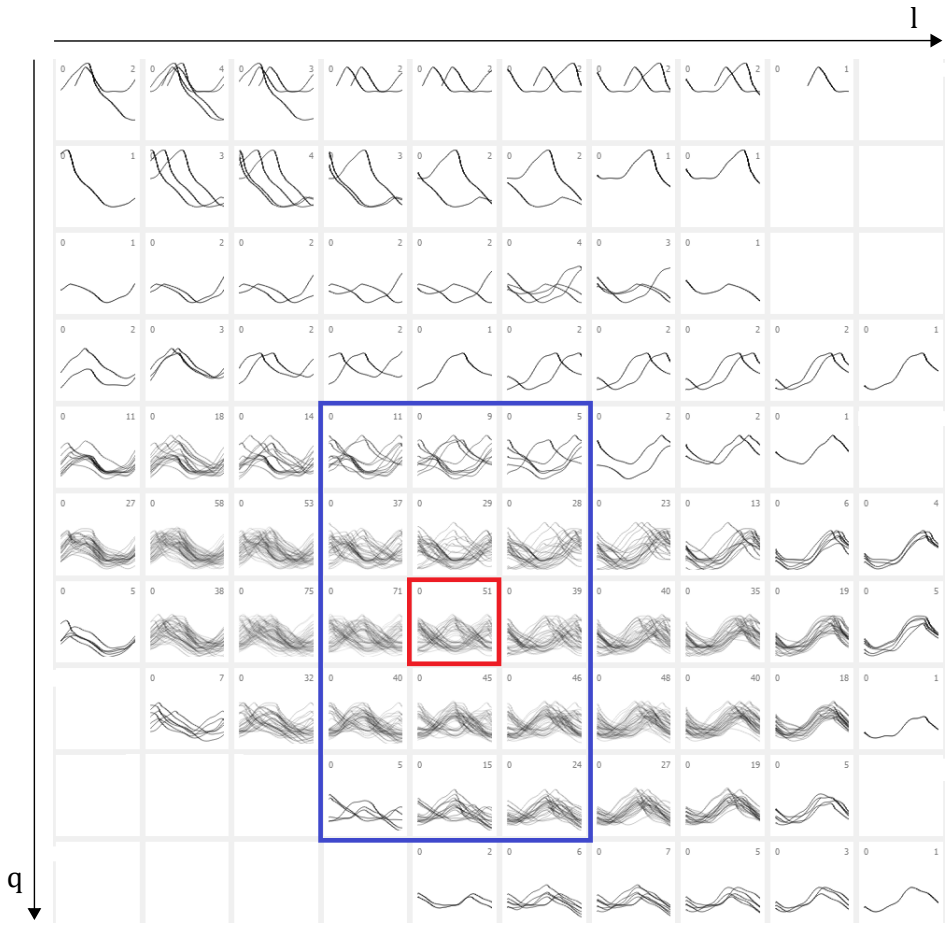



**Figure 52:** *Observing the behavior of* d *for different configurations of* l *and* q*. The variance in the critical region (blue rectangle) is characterized by two sets of diverging curves. Those easiest to separate (red rectangle) are used for further analysis of the target variance.*

## 7.2 TARGET-ORIENTED EXPLORATION OF FEATURE SUBSETS

In this second part of the Proof of Concept, we perform the first step of a feature selection procedure based on l and q as independent features and d as the target feature to be predicted. We start by analyzing the Synchronization Grid to assess how well the feature subset {l, q} explains d as the target feature.

### 7.2.1 *Gaining an Overview Using the Synchronization Grid*

The Synchronization Grid is shown in Figure 52. Its upper third contains Synchronization Line Plots depicting only a few curves (i.e. up to four). This is caused by the sparse coverage of the feature space in this area. For reasons of missing significance, we will not consider this region for further analysis. A certain variance is present for the cells in the left and right thirds of the grid, but the slope of the curves can be said to be consistent for these cells. They either show decreasing (left third) or increasing (right third) target values starting from the synchronization point. Due to this consistency, we do not consider them as critical.

Instead, we will focus on those parts of the feature space, for which the target feature shows significantly discordant behavior (Figure 52, blue rectangle). In this region, both decreasing and increasing target values occur as behaviors. Consequently, the corresponding cells mainly contain two sets of diverging curves. This leads to the conclusion that there must be an additional influencing feature, which discriminates between both sets and thus determines the behavior that the target actually shows. The question is: based on which feature can we decide whether the target behavior belongs to one or the other set of curves? In other words, which feature can be used to discriminate between a target increase and a decrease? Having found such a feature, adding it to the set of predictors might significantly reduce the target variance and thus improve the explanatory power. However, such conclusions need to be double-checked by assessing the target variance for the newly created feature set.

### 7.2.2 *Drill-Down to the Region of Interest*

We now investigate one of the cells from the region, to which the analysis was already narrowed down, in more detail to identify the next feature to be added to the feature subset. A suitable feature shows a significant difference in the distributions associated with one of the sets of diverging curves and the entirety of curves. Data items that belong to both sets of curves make it difficult to perceive a difference between the distributions, because they contribute to both of them. If many of them are present, a comparison might be inconclusive. Accordingly, we choose the cell (Figure 52, red rectangle), for which the sets of diverging curves are best separable (Figure 53a). In this way, we prevent an overlapping of the two sets (i.e. green set against all curves) to be compared in the next step.

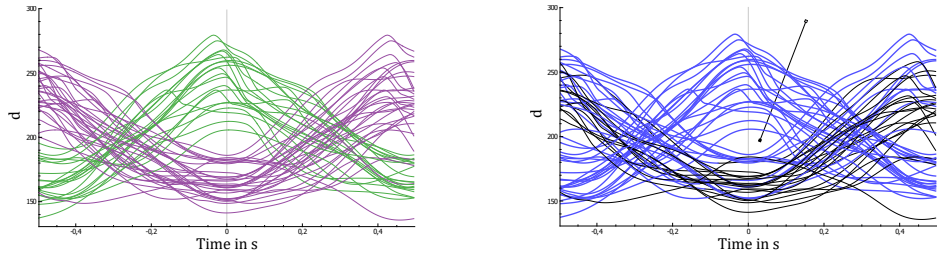
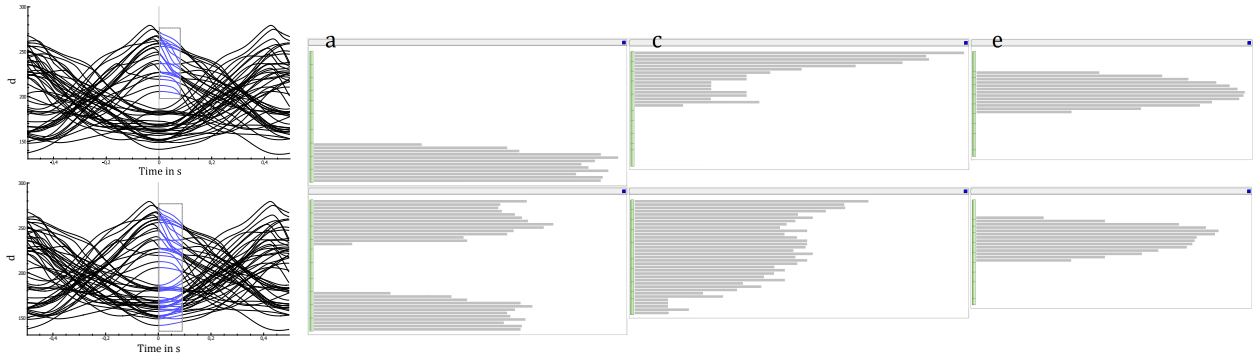



**Figure 53:** *Critical cells contain two sets of curves with contrary behavior that make up the variance (a). Comparing one set to all curves might reveal a feature that relates to the variance. However, undesired curves are included in a line brush due to replication of curve sections (b).*

We first choose the line brush to easily select one set without including curves from the other set. Due to the highly frequent time series in this data set, the configuration trajectory crosses many different grid cells in a short time. Consequently, the automatically selected synchronization points exhibit a small temporal distance. Together with the fixed interval length, this leads to strong replications. Brushing only one of the two diverging sets leads to the majority of curves – 40 out of 51 – being selected (Figure 53b). Obviously, this is not effective for comparing one of the sets to all curves. For this reason, we change the mode to a rectangular brush. As it allows us to limit the brush to a certain time range, we can specify those parts, in which the curves actually diverge, more concretely. At the same time, covering only sections of the curves instead of the whole reduces the selection of replicated curves belonging to the other set. Thus, it offers a better separation of the two brushes.

### 7.2.3 *Comparing Focus and Context Distributions*

Having specified the focus and context brushes (Figure 54a), we can now compare the corresponding distributions for all remaining features. We notice differences for nearly all features – for some to a greater and for others to a lesser extent. In all cases, the focus brush can be associated with values covering a rather small part of the feature domain. Only b, e, and k can be excluded as candidates, because their distributions do not change when brushing the region of interest instead of all curves (Figure 54b). The remaining features have to be compared carefully to identify those with the most significant differences as predictor candidates.




**(a)** *Focus + Context* **(b)** *Corresponding Value Distributions*

**Figure 54:** *Comparing the value distributions w.r.t. the focus and context brushes (a). Feature* e *is excluded as predictor, because the distributions are too similar (b). Brushing the variance significantly influences the distribution of* a*, indicating an additional predictor.*

For two features, namely a and j, the context distribution shows two peaks to each end of the feature domain, while the brush distribution only covers one of these peaks (see Figure 54b). Thus, brushing the region of interest obviously has a significant influence on the distribution. We can even go a step further and compare the distribution of the brushed region with that of its complement. The histograms reveal that the complement's distribution is also the complement, i.e. the second peak, of the context distribution. Thus, the values of these features are well-suited to discriminate between both sets of diverging curves in the Synchronization Line Plot. As a was already identified as an independent feature before, we solely consider a as a predictor candidate.

Observing the distributions of the remaining features, it cannot be clearly determined which of them should additionally be analyzed in more detail as a predictor candidate. However, we can distinguish between two groups of features with different characteristics. 1) For some features (e.g. f or h), the focus distribution's shape is highly similar to a specific part of the context distribution – as if the context distribution was cut in two parts, where one of those parts is the focus distribution (Figure 54b). In this sense, the difference between focus and context lies in the covered part of the feature domain, but not in the distributions' shapes. 2) For other features (e.g. c, m, n, or o), the focus distribution has a completely different shape than the context distribution within the same range (Figure 54b). Here, the difference between focus and context distribution does

not only originate from the difference in the covered ranges of the feature, but also in the significantly different shapes of (part of) the distributions. As a conclusion, besides a, we cannot decide for another potential predictor to be added to the feature set, because no feature shows a significantly larger difference in its distributions than the other features.

**Finding 0.6 (Target-Oriented Exploration)** *Given* {l, q} *for prediction of* d*, we notice that this feature subset is not sufficient. Features* b*,* e*, and* k *are excluded as predictors. Feature* a *is considered as a predictor candidate. The remaining features cannot be rejected nor confirmed as suitable predictor candidates.*

## 7.3 SUMMARY OF RESULTS AND COMPARISON TO GROUND TRUTH

In this section, we reveal the relationships that were intentionally included in the data set. While doing so, we evaluate to what extent the findings gained with our approach actually match this ground truth. This gives us an impression of the system's suitability for relationship discovery and feature selection in multivariate time series data.

The data set originates from a predator-prey simulation involving plants, rabbits, and foxes. For each population, the number of living and eaten organisms, energy and water level, as well as reproduction are simulated. Two external influencing factors, namely precipitation and light are also included. All three populations depend on the available amount of water. The plants additionally need light as energy source. Due to the complexity of the simulated system, the data set contains various dependencies. Some of them are direct, like the relationship between eaten plants and the number of rabbits. Others involve more independent features, e.g. the reproduction of rabbits depends on both their water and energy level.

### FINDING 0.1: a AND b ARE INDEPENDENT

This finding can be confirmed by the ground truth. The features a and b correspond to *light* and *precipitation*, which were included in the simulation as external factors to bound the maximum size of the populations.

### FINDING 0.2: d = g AND l = r

These perfect dependencies are also contained in the data set. The features d and g represent the *foxes' energy level* and the *number of living rabbits*. For simulation, the energy of the foxes was modeled to equal the number of rabbits that are available as food. The same holds for r and l representing the *rabbits' energy level* and the *number of plants*.

FINDING 0.3: q INFLUENCES d

This can be partially confirmed. Feature q represents the *number of foxes*, while d stands for the *foxes' energy*, which is equal to the *number of rabbits* (g). According to the simulation model, the number of rabbits indeed depends on the number of foxes. We noticed that low values of q occur together with high values of d and vice versa (Section 7.1.2). This corresponds to general reasoning: if there are only a few foxes, the rabbits do not have as much enemies threatening their lives. However, the simulation model also suggests, that q is not the only feature influencing d.

FINDING 0.4: f DEPENDS ON k AND l

This can also be partially confirmed. Feature f represents the *reproduction of rabbits*, while k refers to the *reproduction of foxes* and l to the *number of plants*, which in turn equals the *rabbits' energy* (r). The dependency of the reproduction of rabbits on their energy level is contained in the model. Its dependency on the reproduction of foxes is not as straight-forward. One could reason that an increasing number of foxes decreases the number of rabbits available for reproduction, thus reducing the rabbits' reproduction itself. However, this reasoning relies on indirect dependencies and, instead of k, the simulation model involves another feature influencing f.

FINDING 0.5: n SHOULD BE ADDED TO {k, l} TO EXPLAIN f

Searching for an additional feature influencing f to improve the explanatory power of the feature subset {k, l} resulted in n, the *rabbits' water level*. Indeed, the simulation model consists of n and l as features influencing f. Our identified feature subset {k, l, n} contains the truly relevant features, but is not the minimal descriptive subset {l, n}. The proposed approach is robust w.r.t. the choice of initial features: although k was wrongly chosen, the predictor n was correctly identified in the second attempt.

FINDING 0.6: a SHOULD BE ADDED TO {l, q} TO EXPLAIN d

This cannot be confirmed. The task description "Given l and q, on which feature does d additionally depend?" originated from an actual relationship in the simulation data. It corresponds to the *number of rabbits* (g=d) being influenced by the *number of plants* (l) and the *number of foxes* (q). We identified a, i.e. the *light*, to be the missing feature for prediction. According to the simulation model, the *stored water* (c) would have been directly related to d. The identified dependency on light is therefore an indirect one, as stored water is in turn influenced by light due to evaporation. However, using the proposed approach the analyst did not find the most influencing additional feature.

# 8 CONCLUSION

For the purpose of forecasting, time series analysis and modeling is of fundamental importance in numerous application domains. The challenge lies in filtering the relevant from the available features to build a simple model that accurately predicts the target without suffering from dimensionality issues. Feature selection therefore aims at identifying the minimal subset of features that together are most useful for capturing all characteristics of a target feature. Wrapper methods conventionally measure the quality of a feature subset according to the performance of a given predictor that was optimized based on the respective subset. In contrast, filter approaches do not require a predefined analytical model and are not influenced by the performance of the model fitting algorithm.

This thesis contributes a stand-alone filter approach for feature selection in multivariate time series data as a preprocessing step for a regression analysis. It builds upon visualization and interaction techniques to integrate the analyst's strengths into the feature selection process. To comply with the model-free characteristic, we employ a generic quality criterion that is derived from the assumption that a valid model outputs the same predictions when given the same inputs. This assumption can be verified for a given feature subset by using the proposed synchronization approach. It is adapted from an alignment concept in the health care domain, where analysts investigate how time series are affected by the occurrence of an event. The synchronization supports an observation of the similarity of time series that represent the predictions for a particular combination of input values. A broad evaluation of a feature subset's quality thus boils down to visually assessing the variance in sets of curves that refer to different inputs across the entire feature space.

The key visualization that enables analysts to visually approach the quality of a feature subset is surrounded by a system that provides additional standard visualizations supporting an efficient exploration of relationships. Interaction techniques enable the analyst to focus on relevant parts of the data and to steer the feature selection by navigating between different perspectives based on analysis tasks and previous findings. Analysts can also integrate domain knowledge by interactively initializing and re-

fining feature subsets based on their expertise. An informal evaluation was performed by applying the developed method to real-world sensor data as well as an artificial data set. Domain knowledge in the real-world case study was used to initialize the feature subset. The Synchronization Grid offered a good overview of the explanatory power and drew the analyst's attention to critical regions, which served as a starting point for interactively searching for an additional influencing factor. For the artificial data set, the surrounding Interactive Visual Analysis system provided good support in identifying initial features based on relationships. Even if the choice of initial features is not optimal, the true explanatory features are identified as influencing factors in the further process of the analysis.

The presented approach is independent of a model class and not tuned to the model generating process. This property offers several advantages: (1) the function class of the model does not have to be known, (2) low computational effort is required, and (3) due to its independence, it can be combined with any model generating algorithm. Using the strengths of the human visual sense for evaluation of feature subsets makes the feature selection procedure much more tangible than could be achieved by quantifying the quality of a feature subset. On the other hand, analysts face a certain learning effort to be able to efficiently interpret the underlying Synchronization Line Plots. Once the interpretation is clear, the analysis process becomes more transparent, thus allowing for a deeper understanding and interpretation of obtained findings. A strong limitation, however, lies in the low dimensionality: the approach can only be used to evaluate two-dimensional feature subsets and to identify a third predictor to be added. Consequently, only one iteration of the feature selection process can be performed. An evaluation of feature subsets containing three or more features requires an adaptation of the visualization.

The proposed method is intended to serve as a universal dimension reduction step prior to regression modeling. Nominal features are not supported. The approach does not prefer or defer any type of relationship and minimizes the risk of rejecting a relevant feature, which might have resulted in a better feature subset. Because it is based on an Interactive Visual Analysis concept, the system might also be combined with visualization solutions dealing with other steps of the regression pipeline, e.g. the actual model building. The presented method is highly generic. In theory, it is therefore applicable to any modeling task involving multivariate, numerical time series. A more formal and diverse evaluation is needed to demonstrate the method's usefulness for further application scenarios.

# 9 FUTURE WORK

There are several ways to extend the methodology presented in this thesis. In the following, we cover selected ideas targeted at different aspects of an interactive feature selection that might advance the proposed approach.

The first suggestions deal with the quality of the underlying data. Up to now, no particular actions were taken to ensure good data quality. Missing data in certain features, e.g. represented by a *not available* entry, might induce artificial dependencies, due to which relevant features might be erroneously excluded from the feature subset. Noise might also distort the results of modeling, because the data does not accurately represent the underlying real-world conditions. Such errors in the values of features should consequently already be considered for feature selection. In contrast to noise, outliers include not only errors, but might also contain useful information that originate from natural variations within measurements. To explicitly consider missing data, noise, and outliers might yield improved feature selection results that lead to more accurate models.

Grouping the configurations for an overview visualization was performed in a straight-forward way. Defining similar configurations to be located within one grid cell is intuitive, but might also be inappropriate, because it does not consider the distribution of data points within the feature space. More advanced methods like circular neighborhoods or clustering algorithms might reduce the deviation of data points in the same neighborhood and thus improve the precision of the synchronization approach.

To reduce the cognitive load of the analyst, user guidance could be enhanced. For the Synchronization Grid, this might be implemented as an automatic highlighting of critical cells with large target variance to more clearly draw the analyst's attention towards parts of the feature space that need refinement. Guidance might also be realized by initially providing a reasonable choice of parameters, like the grid resolution, to enable users to get started with the analysis right away. Both approaches might also be combined with a guided search strategy to more efficiently identify the most critical parts of the feature space.

Sometimes, a target feature might not directly depend on an original feature, but rather on the way its values change, i.e. the first-order derivation. In the Synchronization Grid, changes of a feature are represented by individual rows or columns. To observe how certain patterns are reproduced throughout the grid could offer meaningful insights concerning the relationship between target variance and changes of features. Thus, advanced interpretation and realization of brushing within the Synchronization Grid might provide answers regarding the question whether first-order derivations might be suitable explanatory features.

Finally, the presented method in this thesis might not only serve as a preprocessing step for model building, but could also be used for model validation. The synchronization approach was developed in a general manner: it simply takes a feature subset as input and does not require further assumptions about the data or future model. In theory, it cannot only be applied to training data for feature selection, but also to a time series that was predicted by an already fitted model, which allows for an evaluation of the model's explanatory power. However, further evaluation of the method's usage as a model validation approach is required.

# BIBLIOGRAPHY


[1] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. "Visualizing Time-Oriented Data—A Systematic View." In: *Computers & Graphics* 31.3 (2007), pp. 401–409.

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. Springer-Verlag GmbH, June 11, 2011. ISBN: 0857290789.

[3] Hussein Almuallim and Thomas G Dietterich. "Learning With Many Irrelevant Features." In: *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 91. 1991, pp. 547–552.

[4] Robert Amar and John Stasko. "A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations." In: *IEEE Symposium on Information Visualization*. IEEE, 2004, pp. 143–150.

[5] J Scott Armstrong. "Illusions in Regression Analysis." In: (2011).

[6] Michael A Babyak. "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models." In: *Psychosomatic Medicine* 66.3 (2004), pp. 411–421.

[7] 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." In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 559–568.

[8] Richard Bellman. *Dynamic Programming*. Courier Corporation, 2013.

[9] David A Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, 1991.

[10] Richard A Berk. *Regression Analysis: A Constructive Critique*. Vol. 11. Sage, 2004.

[11] Mairead L Bermingham, Ricardo Pong-Wong, Athina Spiliopoulou, Caroline Hayward, Igor Rudan, Harry Campbell, Alan F Wright, James F Wilson, Felix Agakov, Pau Navarro, et al. "Application of High-Dimensional Feature Selection: Evaluation for Genomic Prediction in Man." In: *Scientific Reports* (2015).

[12] Gavin Brown. "A New Perspective for Information Theoretic Feature Selection." In: *Artificial Intelligence and Statistics*. 2009, pp. 49–56.

[13] 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 International Conference on Information Visualization (IV)*. IEEE. 2007, pp. 191–196.

[14] Claire Cardie. "Using Decision Trees to Improve Case-Based Learning." In: *Proceedings of the Tenth International Conference on Machine Learning*. 1993, pp. 25–32.

[15] Samprit Chatterjee and Ali S Hadi. *Regression Analysis by Example*. John Wiley & Sons, 2015.

[16] William C Cleveland and Marylyn E McGill. *Dynamic Graphics for Statistics*. CRC Press, Inc., 1988.

[17] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, Aug. 1, 2006. 236 pp. ISBN: 0521685672.

[18] David R Cox and Nanny Wermuth. "Some Statistical Aspects of Causality." In: *European Sociological Review* 17.1 (2001), pp. 65–74.

[19] Selan Dos Santos and Ken Brodlie. "Gaining Understanding of Multivariate and Multidimensional Data Through Visualization." In: *Computers & Graphics* 28.3 (2004), pp. 311–325.

[20] Lynn E Eberly. "Topics in Biostatistics." In: ed. by Walter T Ambrosius. Humana Press, 2007. Chap. Correlation and Simple Linear Regression, pp. 143–164.

[21] H Gray Funkhouser. "A Note on a Tenth Century Graph." In: *Osiris* 1 (1936), pp. 260–262.

[22] Donna L Gresh, Bernice E Rogowitz, Raimond L Winslow, David F Scollan, and Christina K Yung. "WEAVE: A System for Visually Linking 3-D and Statistical Visualizations, Applied to Cardiac Simulation and Measurement Data." In: *Proceedings of the Conference on Visualization*. IEEE Computer Society Press. 2000, pp. 489–492.

[23] 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)*. IEEE. 2011, pp. 43–50.

[24] Diansheng Guo. "Coordinating Computational and Visual Approaches for Interactive Feature Selection and Multivariate Clustering." In: *Information Visualization* 2.4 (2003), pp. 232–246.

[25] Zhenyu Guo, Matthew O. Ward, and Elke A. Rundensteiner. "Model Space Visualization for Multivariate Linear Trend Discovery." In: *2009 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009.

[26] Isabelle Guyon and André Elisseeff. "An introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3.Mar (2003), pp. 1157–1182.

[27] Robin Haring, Henri Wallaschofski, Matthias Nauck, Marcus Dörr, Sebastian E Baumeister, and Henry Völzke. "Ultrasonographic Hepatic Steatosis Increases Prediction of Mortality Risk from Elevated Serum Gamma-Glutamyl Transpeptidase Levels." In: *Hepatology* 50.5 (2009), pp. 1403–1411.

[28] Helwig Hauser. "Generalizing Focus+Context Visualization." In: *Scientific Visualization: The Visual Extraction of Knowledge from Data*. Springer, 2006, pp. 305–327.

[29] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. "Angular Brushing of Extended Parallel Coordinates." In: *IEEE Symposium on Information Visualization (InfoVis)*. 2002, pp. 127–130.

[30] Susan Havre, Beth Hetzler, and Lucy Nowell. "ThemeRiver: Visualizing Theme Changes over Time." In: *IEEE Symposium on Information Visualization (InfoVis)*. 2000, pp. 115–123.

[31] Thomas Herndon, Michael Ash, and Robert Pollin. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." In: *Cambridge Journal of Economics* 38.2 (2014), pp. 257–279.

[32] Harry Hochheiser and Ben Shneiderman. "Interactive Exploration of Time Series Data." In: *Discovery Science*. Springer. 2001, pp. 441–446.

[33] Harry Hochheiser and Ben Shneiderman. "Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration." In: *Information Visualization* 3.1 (2004), pp. 1–18.

[34] Gordon Hughes. "On the Mean Accuracy of Statistical Pattern Recognizers." In: *IEEE Transactions on Information Theory* 14.1 (1968), pp. 55–63.

[35] Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Moller. "DimStiller: Workflows for Dimensional Analysis and Reduction." In: *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2010.

[36] Alfred Inselberg and Bernard Dimsdale. "Parallel Coordinates for Visualizing Multi-Dimensional Geometry." In: *Computer Graphics 1987*. Springer, 1987, pp. 25–44.

[37] Waqas Javed, Bryan McDonnel, and Niklas Elmqvist. "Graphical Perception of Multiple Time Series." In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 927–934.

[38] George H John, Ron Kohavi, Karl Pfleger, et al. "Irrelevant Features and the Subset Selection Problem." In: *Machine Learning: Proceedings of the Eleventh International Conference*. 1994, pp. 121–129.

[39] Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

[40] Daniel A Keim. "Visual Exploration of Large Data Sets." In: *Communications of the ACM* 44.8 (2001), pp. 38–44.

[41] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. "Visual Analytics: Definition, Process, and Challenges." In: *Information Visualization*. Springer, 2008, pp. 154–175.

[42] Kenji Kira and Larry A Rendell. "A Practical Approach to Feature Selection." In: *Proceedings of the Ninth International Workshop on Machine Learning*. 1992, pp. 249–256.

[43] Willi Klösgen and Jan M Zytkow. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, Inc., 2002.

[44] Ron Kohavi and George H John. "Wrappers for Feature Subset Selection." In: *Artificial intelligence* 97.1-2 (1997), pp. 273–324.

[45] Zoltan Konyha, Kresimir Matkovic, Denis Gracanin, Mario Jelovic, and Helwig Hauser. "Interactive Visual Analysis of Families of Function Graphs." In: *IEEE Transactions on Visualization and Computer Graphics* 12.6 (2006), pp. 1373–1385.

[46] Samuel Kotz, ed. *Encyclopedia of Statistical Sciences*. Vol. 2. John Wiley & Sons, 2006.

[47] Samuel Kotz, ed. *Encyclopedia of Statistical Sciences*. Vol. 3. John Wiley & Sons, 2006.

[48] Josua Krause, Adam Perer, and Enrico Bertini. "INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data." In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1614–1623.

[49] 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.

[50] Wojtek Krzanowski. *Principles of Multivariate Analysis*. OUP Oxford, 2000.

[51] Debbie A Lawlor, George Davey Smith, and Shah Ebrahim. "Commentary: The Hormone Replacement - Coronary Heart Disease Conundrum: Is this the Death of Observational Epidemiology?" In: *International Journal of Epidemiology* 33.3 (2004), pp. 464–467.

[52] Erich Leo Lehmann et al. "Some Concepts of Dependence." In: *The Annals of Mathematical Statistics* 37.5 (1966), pp. 1137–1153.

[53] Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. *Forecasting: Methods and Applications*. John Wiley & Sons, 2008.

[54] Kresimir Matkovic, Wolfgang Freiler, Denis Gracanin, and Helwig Hauser. "ComVis: A Coordinated Multiple Views System for Prototyping New Visualization Technology." In: *12th International Conference on Information Visualisation (IV)*. IEEE. 2008, pp. 215–220.

[55] Thorsten May, James Davey, and Tobias Ruppert. "Smartstripes - Looking Under the Hood of Feature Subset Selection Methods." In: *Proceedings of the 2nd International Workshop on Visual Analytics (EuroVA)*. 2011, pp. 13–16.

[56] Thorsten May, Andreas Bannach, James Davey, Tobias Ruppert, and Jörn Kohlhammer. "Guiding Feature Subset Selection with an Interactive Visualization." In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2011.

[57] Thomas Mühlbacher and Harald Piringer. "A Partition-Based Framework for Building and Validating Regression Models." In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 1962–1971.

[58] Philipp Muigg, Johannes Kehrer, Steffen Oeltze, Harald Piringer, Helmut Doleisch, Bernhard Preim, and Helwig Hauser. "A Four-level Focus+ Context Approach to Interactive Visual Analysis of Temporal Features in Large Scientific Data." In: *Computer Graphics Forum*. Vol. 27. 3. Wiley Online Library. 2008, pp. 775–782.

[59] Amnon Naamad, DT Lee, and W-L Hsu. "On the Maximum Empty Rectangle Problem." In: *Discrete Applied Mathematics* 8.3 (1984), pp. 267–277.

[60] Chris North and Ben Shneiderman. "Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata." In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM. 2000, pp. 128–135.

[61] Steven Novella. "Evidence in Medicine: Correlation and Causation." In: *Science-Based Medicine* 18 (2009).

[62] Jane A Ou and Stephen H Penman. "Financial Statement Analysis and the Prediction of Stock Returns." In: *Journal of Accounting and Economics* 11.4 (1989), pp. 295–329.

[63] Hans Pacejka. *Tire and Vehicle Dynamics*. Elsevier, 2005.

[64] Alan Pankratz. *Forecasting with Dynamic Regression Models*. Vol. 935. John Wiley & Sons, 2012.

[65] Harald Piringer, Wolfgang Berger, and Helwig Hauser. "Quantifying and Comparing Features in High-Dimensional Datasets." In: *International Conference Information Visualisation (IV)*. IEEE, 2008.

[66] Hannes Reijner et al. "The Development of the Horizon Graph." In: (2008).

[67] Jonathan C Roberts. "State of the Art: Coordinated & Multiple Views in Exploratory Visualization." In: *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV)*. IEEE. 2007, pp. 61–71.

[68] Daniel Sarewitz, Roger A Pielke, and Radford Byerly. *Prediction: Science, Decision Making, and the Future of Nature*. Island Press, 2000. ISBN: 978-1559637763.

[69] Dieter Schramm, Manfred Hiller, and Roberto Bardini. *Vehicle Dynamics: Modeling and Simulation*. Springer, 2014. ISBN: 978-3-540-36044-5.

[70] Jinwook Seo and Ben Shneiderman. "A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data." In: *Information Visualization* 4.2 (2005), pp. 96–113.

[71] Ben Shneiderman. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." In: *Proceedings of the IEEE Symposium on Visual Languages*. IEEE. 1996, pp. 336–343.

[72] Emre Soyer and Robin M Hogarth. "The Illusion of Predictability: How Regression Statistics Mislead Experts." In: *International Journal of Forecasting* 28.3 (2012), pp. 695–711.

[73] Christian Tominski, James Abello, and Heidrun Schumann. "Axes-Based Visualizations with Radial Layouts." In: *Proceedings of the ACM Symposium on Applied Computing*. ACM. 2004, pp. 1242–1247.

[74] Edward R Tufte. *The Visual Display of Quantitative Information*. 1983.

[75] Edward R Tufte. *The Cognitive Style of PowerPoint*. Vol. 2006. Graphics Press Cheshire, CT, 2003.

[76] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, 2006. ISBN: 0961392177.

[77] Michael Unterreiner. "Modellbildung und Simulation von Fahrzeugmodellen unterschiedlicher Komplexität." PhD thesis. Universität Duisburg-Essen, Fakultät für Ingenieurwissenschaften, 2013.

[78] Michael Unterreiner and Dieter Schramm. "Modelling of a Twin-Track Vehicle Model with Modular Wheel Suspensions." In: *Applied Mechanics and Materials*. Vol. 165. Trans Tech Publ. 2012, pp. 214–218.

[79] Erik Verlinde. "On the Origin of Gravity and the Laws of Newton." In: *Journal of High Energy Physics* 2011.4 (2011), p. 29.

[80] Tyler Vigen. *Spurious Correlations*. Hachette Books, 2015.

[81] Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, Dörte Radke, Roberto Lorbeer, Nele Friedrich, Nicole Aumann, Katharina Lau, Michael Piontek, Gabriele Born, et al. "Cohort Profile: the Study of Health in Pomerania." In: *International Journal of Epidemiology* 40.2 (2010), pp. 294–307.

[82] Karla Zadnik, Lisa A Jones, Brett C Irvin, Robert N Kleinstein, Ruth E Manny, Julie A Shin, and Donald O Mutti. "Vision: Myopia and Ambient Night-Time Lighting." In: *Nature* 404.6774 (2000), p. 143.

[83] Jian Zhao, Fanny Chevalier, Emmanuel Pietriga, and Ravin Balakrishnan. "Exploratory Analysis of Time-Series with Chronolenses." In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2422–2431.

[84] Kelly H Zou, Kemal Tuncali, and Stuart G Silverman. "Correlation and Simple Linear Regression." In: *Radiology* 227.3 (2003), pp. 617–628.