Significance of Patterns in Data Visualisations

Rafael Savvides
Andreas Henelius
Emilia Oikarinen
Kai Puolamäki
rafael.savvides@helsinki.fi
andreas.henelius@helsinki.fi
emilia.oikarinen@helsinki.fi
kai.puolamaki@helsinki.fi
Department of Computer Science
University of Helsinki
Helsinki, Finland

ABSTRACT

In this paper we consider the following important problem: when we explore data visually and observe patterns, how can we determine their statistical significance? Patterns observed in exploratory analysis are traditionally met with scepticism, since the hypotheses are formulated while viewing the data, rather than before doing so. In contrast to this belief, we show that it is, in fact, possible to evaluate the significance of patterns also during exploratory analysis, and that the knowledge of the analyst can be leveraged to improve statistical power by reducing the amount of simultaneous comparisons. We develop a principled framework for determining the statistical significance of visually observed patterns. Furthermore, we show how the significance of visual patterns observed during iterative data exploration can be determined. We perform an empirical investigation on real and synthetic tabular data and time series, using different test statistics and methods for generating surrogate data. We conclude that the proposed framework allows determining the significance of visual patterns during exploratory analysis.

CCS CONCEPTS

• **Mathematics of computing** → **Exploratory data analysis**; *Nonparametric statistics*; *Multivariate statistics*; Time series analysis; • **Human-centered computing** → *Visual analytics.*

KEYWORDS

visual analytics; significance testing; exploratory data analysis

ACM Reference Format:
Rafael Savvides, Andreas Henelius, Emilia Oikarinen, and Kai Puolamäki. 2019. Significance of Patterns in Data Visualisations. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330994

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3330994

1 INTRODUCTION

A fundamental question in exploratory data analysis (EDA), especially when the amount of data is limited, concerns whether patterns visually observed by a human analyst are real or just random artefacts.

In the typical frequentist approach this problem is given, e.g., by the statistical significance testing method, where the expert is supposed to formulate the hypotheses prior to looking at the data. This, however, does not correspond to practice: typically, the analyst explores the data and formulates hypotheses *during data exploration*. The naïve approach to this problem is to formulate all possible hypotheses prior to investigating the data. To do this properly, one would hence have to control for the effect of multiple hypotheses being tested. Because the number of potential hypotheses is typically very large, it means that all statistical power is easily lost. Therefore, one would (naïvely) assume that statistical significance and visual data exploration mix badly with each other. In this paper we argue that it is, in fact, possible to do exploration and at the same time find statistically meaningful visual patterns.

Although the results of EDA are not used for inference, it is inevitable that any exploration influences the analyst's perception of the data and instils some bias in later analyses. However, EDA can be combined with expert knowledge to discover interesting patterns. When searching for relationships between variables, non-experts might redundantly search for irrelevant relationships, i.e. test a large number of irrelevant hypotheses. In contrast, domain experts ask focused and specific questions, making them more likely to search for relevant relationships, thereby reducing the number of required hypotheses and increasing statistical power.

The problem of over-interpreting data is especially important in iterative exploration scenarios, such as projection pursuit, since visualising multiple views of the same data inevitably results in some discovered patterns, i.e., the multiple comparisons problem in visualisation [33]. We now present two simple examples motivating our approach.

Motivating Example 1. As our first example, we use the time series data shown in Fig. 1, representing the hourly carbon monoxide (CO) concentration for a single day (March 25th, 2004) from the UCI [7] `Air quality` dataset [6]. A typical question an analyst might ask is *whether the values of the time series at some given time*

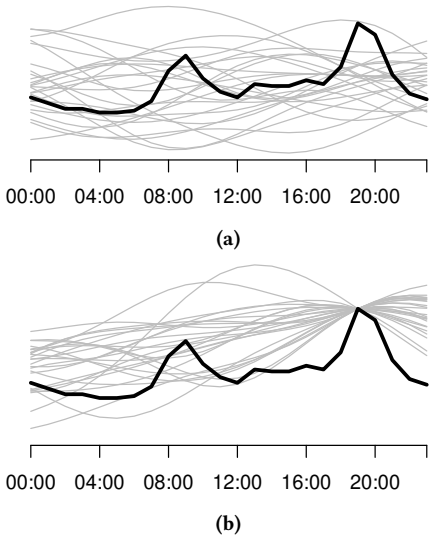


Figure 1: The hourly carbon monoxide (CO) concentration for one day (the black line). The gray lines are Gaussian process priors in (a) and posteriors in (b). (a) There are spikes in the morning and the evening due to people commuting to and from work. Using the framework, we find that the afternoon spike is significant. As a test statistic we use the value of the time series at each point and as a null model we use Gaussian processes with squared exponential kernel. The minP adjusted p-value for the afternoon spike at 19:00 is 0.1. (b) When we constrain the null model at the maximum point, there are no longer values that are significantly larger than the values in the null model.

instances are extreme compared to his or her expectations. The problem can be formulated as a statistical hypothesis testing problem as follows. We define a separate test statistic for each time point in the time series, the value of the test statistic being the value of the time series at that time point. Additionally, we need a null distribution of time series that reflects the analyst's assumptions of the behaviour of the time series before observing anything. Here, we take this distribution to be the prior of a Gaussian process with a squared exponential kernel. Because there are many hypotheses (each test statistic corresponds to a particular hypothesis) we must use multiple comparisons correction, such as the minP test [28]. After this correction, we notice that only one of the peak values (at 19:00) is extreme using a threshold of $\alpha = 0.1$, as shown in Fig. 1a.

Even though this example is a straightforward implementation of statistical hypothesis testing when samples of the null distribution are available, it shows that these techniques are feasible also for visual patterns during visual exploration. In Sec. 2.2 we show further examples of visual patterns for both time series and tabular data.

Furthermore, we can use the observed statistically significant visual patterns to constrain the null distribution. Fig. 1b shows the updated null distribution, into which we have incorporated the constraint concerning the previously found significantly large value. Under this new null distribution there are no longer time instances where the value of the time series is significantly larger that the values in the null distribution.

Motivating Example 2. Our second example demonstrates exploration of tabular data. We use an extract of the UCI Adult data set [14]. We sample 90 transactions at random from the dataset to simulate a scenario with a limited amount of data. Assume that the analyst is interested in factors correlated with the income level (low income vs. high income). The analyst uses his or her prior knowledge to hypothesise that sex, education level (high vs. low), and marital status (married or not) likely affect the correlation.[1] The data is shown in Tab. 1. By using a simple test statistic (count of items) we compute the adjusted p-values and notice that being married indeed is a significant factor contributing to high income.

We can then modify the null distribution (as in the time series example above) so that the relation between marriage and income is incorporated in the null distribution, after which the null distribution and the data no longer differ in this relation. Using this null distribution we obtain the p-values denoted by $p_{\text{mar. fix.}}$ in Tab. 1 After this we no longer find any significant values and we can conclude that being married explains the high income sufficiently well in this small dataset. Note that if we had had a larger sample from the dataset we would have found also other significant correlations.

Related work. This work is related to visual analytics [13], visual data exploration [19], and graphical (or visual) inference [3, 16, 30, 31]. The framework proposed in this paper can be used to assess the significance of visual patterns. It empowers data analysts by allowing direct investigation of various hypotheses during visual data exploration and is hence relevant to practically any domain involving data analysis and visualisation. There are numerous studies on statistical testing of patterns in data; to the best of our knowledge there are none that have formalised a *general procedure* for testing the *significance of visual patterns*.

Graphical inference [3] formalises visual patterns as test statistics and the discovery of a pattern as a rejection of a null hypothesis. The statistical test is the user's cognition: the user is presented with $p - 1$ plots of simulated data and one plot of the real data. If the

[1] Here we consider a subject to be of high education if his or her *education* attribute is *Bachelors*, *Doctorate*, *Masters*, or *Prof-school*, and married if the *relationship* attribute is *Husband* or *Wife*.

Table 1: Example with an extract of the UCI adult data. The columns show the level of income cross-tabulated against sex, level of education, and marital status for subjects in the dataset. p_{adj} are p-values after adjustment for multiple comparisons and $p_{\text{mar. fix.}}$ are p-values using a null distribution in which the relation between marriage and income is fixed.

	Sex		Education		Married	
Income	**female**	**male**	**low**	**high**	**no**	**yes**
low	32	41	62	11	53	20
high	3	14	11	6	1	16
$\mathbf{p_{adj}}$	1.00	0.12	1.00	0.15	1.00	0.001
$\mathbf{p_{mar.\,fix.}}$	1.00	0.13	1.00	0.13	1.00	1.00

user correctly identifies the plot of the real data from the other plots and explains which feature distinguishes it, then that feature is statistically significant at level $1/p$.

Our approach is similar but the user is only presented with a plot of the real data and the statistical test is quantitative. The visual pattern to be tested is specified beforehand according to the user's knowledge and is explicitly quantified as a function of the plot which measures the strength of the pattern. Visual features have been previously described through score functions; for instance scagnostics [32] describe the global features of a scatterplot and have been applied in the automatic sorting of scatterplots and multivariate time series [1, 4] as well as serving as a projection pursuit index [29]. Local visual features have also been described through motif-based measures [23]. The visualisation community has dealt with uncertainty in visualisations [17] and errors stemming from projecting multivariate data to two-dimensional plots [5, 21, 24]. It has been noted that there is a need for assessing visual discoveries during exploration. For example, a user study [33] found that over half of the user insights obtained by their visualisations were false due to the effect of multiple comparisons.

Statistical significance testing has been used to find most informative set of patterns in non-interactive settings [15], where the authors looked for the most significant set of patterns given one global test statistic.

Summary of contributions. (i) We present a novel framework that allows the statistical significance of visually observed patterns to be evaluated in a principled manner, (ii) we show how visual data exploration augmented by the analyst's knowledge can improve the statistical power by lowering the number of hypotheses that must be simultaneously tested, (iii) we empirically demonstrate the framework by applying it in the analysis of real-world datasets using both tabular data and time series.

Organisation of this paper. In Sec. 2 we first define visual patterns and significance, after which we present the framework for evaluating the significance of visual patterns. In Sec. 3 we empirically demonstrate how the framework works, after which we conclude with a discussion in Sec. 4.

2 METHODS

Our objective is to define a statistical significance testing procedure using which we can give an upper bound for the probability of even one false discovery by a given α, i.e., we want to control the family-wise error rate. In this paper, we always use $\alpha = 0.1$.

The procedure is as follows. We assume that the user iteratively views different visualisations of the data, where we denote the views by $t = 1, 2, \ldots$. We assume that each view contains a distinct set of visual patterns, each of which is associated with a test statistic for which we can compute the p-value. Our control procedure guarantees that if the visual pattern is a random artefact its adjusted p-value is at most α with a probability that is bounded from above by α.

2.1 Patterns and Significance

We adapt the notation from [15]. Let Ω denote the sample space, which includes all possible data samples, and let $\omega_0 \in \Omega$ denote the observed data set. The user's initial background distribution, i.e., the user's prior knowledge of the data, is defined by a probability function Pr over the sample space Ω. We use $\Pr(\omega)$, with $\omega \in \Omega$, to denote the probability of a single data sample ω, and $\Pr(Q)$, where $Q \subseteq \Omega$, to denote the probability mass in Q. $\Pr(Q)$ satisfies $\Pr(Q) = \sum_{\omega \in Q} \Pr(\omega)$.

Let n_T denote the number of pre-defined *test statistics* which correspond to *hypotheses* to be tested. Each test statistic is indexed in $[n_T] = \{1, \ldots, n_T\}$ and is defined by a function $T_i : \Omega \mapsto \mathbb{R}$, where $i \in [n_T]$. Later, we associate each test statistic with a distinct *visual pattern*.

Significance. We assume that the task of the user is to find all test statistics that do not obey the distribution given by $T_i(\omega)$ when $\omega \sim \Pr(\omega)$, and for this purpose we use p-values. The *unadjusted p-value* related to the test statistic $i \in [n_T]$ in an iteration t for a set of constraints is conventionally defined by $p_i^t = \Pr(\Omega_i^+)$, where $\Omega_i^+ = \{\omega \in \Omega \mid T_i(\omega_0) \leq T_i(\omega)\}$, is the probability of observing values of the test statistic at least as high as in the observed data.

Iterative Exploration by Adding Constraints. We assume that for each test statistic there is a *constraint*, a subset of samples $C_i \subseteq \Omega$ which satisfies $\omega_0 \in C_i \subseteq \{\omega \in \Omega \mid T_i(\omega_0) = T_i(\omega)\}$. If the analyst observes the value of a test statistic and it is found significant it makes sense to incorporate this information into the background distribution. Denote by $I \subseteq [n_T]$ a subset of test statistics that the analyst has found significant and by $\Omega_I = \cap_{i \in I} C_i$ with $\Omega_\emptyset = \Omega$. We can update the distribution $\Pr(\omega \mid \Omega_I) \rightarrow \Pr(\omega)$ and repeat the process. Notice that it follows from the definition of the p-value that after updating the distribution the p-values for the test statistics in I satisfy $p_i^t = 1$ for all $i \in I$. Therefore, test statistics that have once been used as constraints can no longer be significant

Multiple Comparisons Correction Within Iterations. If we test more than one hypothesis then some of the unadjusted p-values may become small because of random effects only, which is why we must use *multiple comparisons correction* (MCC) [8]. Typically, the more hypotheses we test, the less powerful the test will be and more likely we are to miss test statistics not obeying the background distribution. Therefore, if the analyst can pick the hypotheses to test in a smart way we can substantially increase our hit rate, or fraction of positives found. We denote the *adjusted p-values* by $\tilde{p}_i^t$.

We use here the minP procedure [28] for multiple comparisons correction. The advantage of the minP test is that it can be computed by using samples from $\Pr(\omega)$ and that it automatically takes into account the correlation structures present in the data, unlike, e.g., Holm-Bonferroni correction. The latter is crucial, as even though there may be a huge number of visual features, the features are often correlated, which makes the effective number of hypotheses smaller and the approach feasible.

MCC Between Iterations. The data mining session consists of "views", each of which gives information of possibly varying sets of test statistics. In principle, if we observe a large enough number of these views we eventually obtain false positives by chance alone. To correct for this we additionally apply a correction procedure to the sequence of views. More specifically, we use the weighted Bonferroni procedure [9] in which we multiply the p-values in each of the views by a factor of $1/w_t$ where t denotes the order of the

view. This is the weighted Bonferroni procedure as long as the weights sum to unity, i.e., $\sum_{t=1}^{\infty} w_t = 1$. In this paper we choose $w_t = 2^{-t}$, which means that we can have an unlimited number of iterations with more statistical power in the first view and the power decreases exponentially as we explore further. In this way we can control the FWER for the whole sequence of iterations at the chosen level. The final adjusted p-value at iteration t is then

$$\tilde{\tilde{p}}_i^t = \min(1, w_t^{-1}\tilde{p}_i^t). \quad (1)$$

Splittable Data. The following observations are important regarding the use of data in our significance testing framework. If the data set is *splittable* into two conditionally independent parts, given the generating model (e.g., i.i.d. data), then the first part can be used to formulate the hypothesis we wish to investigate, while the second part of the data is used for the actual hypothesis testing. However, if the data set is *unsplittable* (as may be the case, e.g., with time series data or network data) we must choose the test statistics to test prior to viewing the data.

2.2 Visual Significance Testing Framework

The significance testing framework follows the typical format of a permutation test, which includes (i) a test statistic and (ii) the distribution of the test statistic under the null hypothesis, to which the observed test statistic is compared. The analysis procedure depends on whether the data is splittable or not. Furthermore, the process can be repeated, if needed.

We next provide an overview of the framework, after which we give examples of test statistics and describe ways of sampling the test statistic under the null hypothesis (i.e., the null distribution).

Overview of the Framework. The analysis proceeds as follows.

(1) At iteration t, show the user a visualisation of the dataset ω_0 containing k_t patterns.
(2) For each pattern, determine the value of the respective test statistic $T_i(\omega_0)$, where $i \in [k_t]$.
(3) Sample R surrogate datasets ω_r, where $r \in [R]$, from the null distribution and calculate the test statistics $T_i(\omega_r)$ for each dataset ω_r and $i \in [k_t]$.
(4) Compute minP MC corrected p-values $\tilde{p}_i^t$ using $T_i(\omega_0)$ and the values $T_i(\omega_r)$.
(5) Adjust the p-values for the current iteration by multiplying the p-values by $w_t^{-1} = 2^t$, giving the final p-values $\tilde{\tilde{p}}_i^t = \min(1,\ w_t^{-1}\tilde{p}_i^t)$, which are deemed significant if $\tilde{\tilde{p}}_i^t \leq \alpha$.
(6) Increment t by one and repeat the process from step 1 until the exploration is over.

Examples of Test Statistics. The test statistic clearly depends on the type of data. As examples we here consider test statistics for tabular data and time series. Note that our goal here is to provide an intuition of test statistics rather than a comprehensive list.

Numerical attributes in *tabular data* can be efficiently visualised using scatterplots, allowing an analyst to observe different features concerning the data. Assume that our hypothesis is that a particular polygonal region R represents a dense region of points, i.e., a cluster. This *hypothesis* concerning region R now corresponds to a particular *test statistic* which can be, e.g., the number of points within R. Note that each hypothesis corresponds to a particular test statistic encapsulating both the hypothesis and the region of interest. More specific descriptors for scatterplots have also been proposed. Based on the work of John and Paul Tukey on *scagnostics* (an abbreviation of *scatterplot diagnostics*), Wilkinson et al. [32] proposed a set of nine numeric scagnostic measures characterising the visual appearance of scatterplots. These measures are used to describe, e.g., how *stringy*, *clumpy*, or *outlying* a scatterplot is.

For time series it is natural to consider hypotheses concerning different time intervals. E.g., let $[t_0, t_1]$ be an interval in the time series from t_0 to t_1 and let the hypothesis be that the time series increases in this interval. A suitable test statistic is then the difference of the value of the time series at t_1 and t_0. As a second example, the test statistic could also be the maximum value in an interval, or the maximum value at a particular time instant (as in Fig. 1).

Null Distribution. The choice of the correct null distribution depends on the data. In general, the distribution of the test statistic under a specific null hypothesis is unknown. In our framework we use the *method of surrogate data* [25] to empirically estimate the sampling distribution of the test statistic under the null hypothesis, i.e., the null distribution. In this method, an ensemble of surrogate datasets consistent with the null hypothesis are generated. The value of the test statistic in the original data is then compared to the values for the ensemble of surrogates, resulting in a p-value.

The methods for *generating surrogate data* can be divided into two main approaches [22, 26]: (i) typical realisations, and (ii) constrained realisations. *Typical realisations* surrogates are obtained by generating a model of the data, e.g., from the original data by autoregressive methods or using Gaussian processes. After the model has been constructed, surrogate samples can be directly obtained from the model. An example of typical realisations is to generate surrogates using a Gaussian distribution with a particular mean and standard deviation, when we want to examine if this is a valid generating model for the observed data. In contrast, the *constrained realisation* surrogates are constructed so that the desired properties are exactly present in the surrogates. One approach to generating constrained surrogates is based on permuting the original data, although in some cases it may be difficult to devise a suitable permutation scheme [27]. An example of constrained realisation surrogates is the use of a Brownian bridge to model stock price time series. Such surrogates represent a random walk on an interval $[a, b]$ constrained such that the value at times a and b is the same for all surrogates.

In addition to generating surrogate data we also consider the case in which we can directly sample datasets from the same distribution as the observed dataset being investigated. We refer to this case as *historic surrogates*. As an example, consider that we investigate the fluctuation in the price change of a particular stock on a particular day. Now, instead of generating surrogates to investigate a hypothesis we can use actual historic data as surrogates from time series representing the price change of the stock during different days. Historic surrogates are an example of splittable surrogates.

3 EXPERIMENTS

In this section we empirically investigate our framework. We first perform a simulation of how the knowledge of the analyst affects

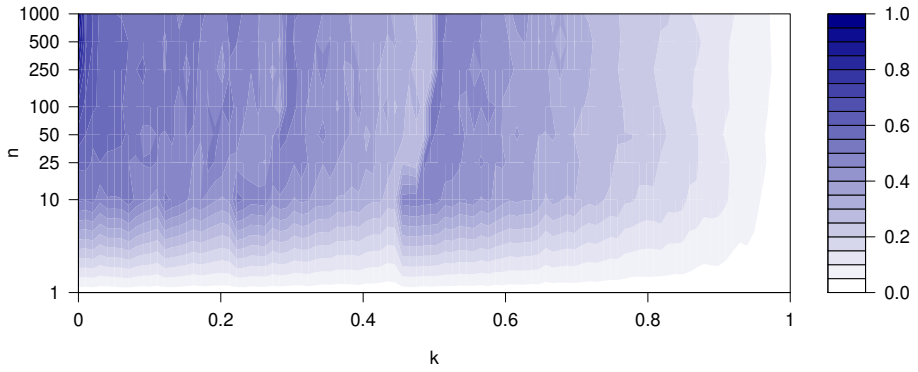


Figure 2: Experiment with synthetic data using $\mu = 3$ and $\beta = 0.5$. The parameter k (on the x-axis) models the experience ("knowledge") of the analyst and n (on the y-axis) the number of test statistics. The colour shows the the mean adjusted p-value of the test statistic for x_1 over 1000 replications.

the the significance of visual patterns. After this we apply the framework in visual testing of hypotheses concerning patterns in *tabular data* and *time series*. In each example, a baseline solution would be to either apply no MCC adjustment or an overly conservative one (e.g. Bonferroni). The experiments were made using R (version 3.5.2) [20]. For the sake of reproducibility, all code used for the experiments is publicly available from https://github.com/edahelsinki/vispa.

Datasets. We use two real-world datasets in the experiments. (1) The `German` socioeconomic dataset [2, 12] is a tabular dataset containing data from 412 German administrative districts. Each district is represented by 46 socioeconomic, political and geographic attributes.[2] We use the same preprocessing as in [18], resulting in 32 real-valued attributes and two class attributes *Type* (Urban, Rural) and *Region* (West, South, East, North). (2) The UCI [7] `Air Quality` dataset [6] contains time series originating from sensors measuring air quality (e.g., carbon monoxide, nitrogen oxides and Benzene). We here consider a time series showing the hourly concentration of carbon monoxide (CO). We remove entries with missing time stamps. In both datasets, we scale the real-valued variables to zero mean and unit variance. In addition, we use synthetic datasets (described in Sec. 3.1) to simulate the analyst's knowledge.

3.1 Leveraging the Analyst's Knowledge

A user with prior knowledge about the data has a high chance of asking the right questions. In this experiment we present a simulated user study in which we apply the framework in a simple scenario. We use synthetic data: n numbers, one of which is different (e.g. x_1). The user has access to a test which can be used to determine if a particular number is different. Using this test, the task of the user is to discover true patterns in the data, i.e. that x_1 is different. An expert user is more likely to correctly select x_1, while a non-expert might need to try (or guess) multiple times before selecting x_1. Since each attempt is a hypothesis test, the resulting p-values need to be adjusted for multiple comparisons. If the user tries too many times, the adjustment causes the test to not be able to determine that x_1 is different, meaning that the user fails to make a true discovery. This experiment demonstrates that (1) experts using the framework are more likely to discover that x_1 is different even when there are many numbers, and (2) non-expert using the framework are also likely to discover x_1 when there are few numbers but fail to do so for increasing n due to the multiple comparisons correction. In a visual exploration scenario, the numbers are replaced by visual patterns. The expert knows which patterns are likely to be significant and can specify which ones to test before looking at the visualisation. This improves statistical power, i.e. the ability to detect true patterns and reduce false negatives. We next look at this experiment in more formal terms.

To demonstrate the effect of the analyst's knowledge on the outcome of the analysis, we devise a simple single-parameter model defining the analyst's level of *expertise*. The idea is that an expert analyst has prior knowledge about the data, and is able to select the relevant test statistics with a high probability. On the other hand, if the analyst has no prior knowledge of the data (non-expert), we assume that the analyst makes a random selection among the test statistics. In our model, the parameter k describes the probability of choosing the correct test statistic (representing the *knowledge* of the analyst), and with probability $1 - k$ the test statistic is chosen uniformly at random among all test statistics.

Furthermore, we assume that the analyst is rational in the sense that a non-expert analyst queries several test statistics in order to increase the probability of finding a significant one, while an expert analyst may only consider a single test statistic. Let β be the confidence level that the analyst wants to acquire. The analyst wants to repeat the choice for the test statistic R times while guaranteeing that the probability of choosing a significant test statistic is at least β (we use $\beta = 0.5$ here). Then R depends on k and on the number of test statistics n as follows:

$$R = \lceil (\log(1-\beta)/(\log(1-1/n) + \log(1-k)) \rceil.$$

In this experiment we use synthetically generated data. The data consists of n real-valued numbers $X_n = \{x_1, \cdots, x_n\}$, where x_1 is sampled from a Gaussian distribution with mean $\mu = 3$ and standard deviation 1, i.e., from $\mathcal{N}(\mu, 1)$. The other x_i's are sampled from $\mathcal{N}(0, 1)$. As the test statistics, we use the value of each data

[2] Available from http://users.ugent.be/~bkang/software/sica/sica.zip

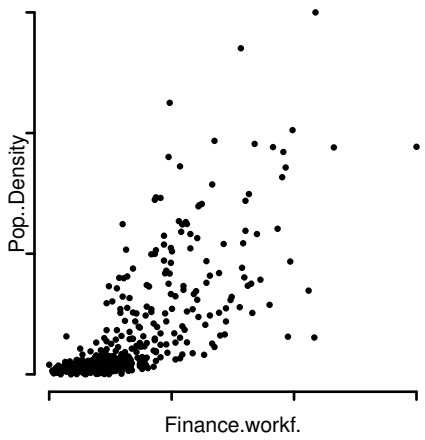


Figure 3: Scatterplot showing attributes `Finance Workforce` and `Population Density` in the `German` data.

item and the null hypothesis is that each value is sampled from $\mathcal{N}(0,1)$. Thus, by the data generation, the assumption is that only the test statistic for x_1 should be significant.

We now determine for different numbers of test statistics n and for different levels of the analyst's knowledge (represented by k) whether the analyst finds the test statistics concerning x_1 to be significant. Depending on the number of times the process of choosing test statistics is repeated (R), the analyst may evaluate several test statistics and minP correction is needed to obtain adjusted p-values (all the test statistics chosen are evaluated in the same step, i.e., there is no iteration step here).

We report the mean adjusted p-value of the test statistic for x_1 over 1000 replications in Fig. 2. If the test statistic for x_1 is not among the test statistics evaluated, the p-value is taken to be 1.00. We observe that when the analyst has a high level of expertise ($k \geq 0.9$), this helps in finding the significant test statistic, regardless of the number of test statistics n. On the other hand, when the analyst has a low level of expertise ($k < 0.2$), then "guessing" the test statistic is likely to fail, even if n is relatively low.

3.2 Testing Hypotheses in Tabular Data

Numerical attributes in tabular data can be efficiently visualised using scatterplots, allowing an analyst to observe different features in the data. In this section we consider the `German` dataset and use (i) scagnostics and (ii) the number of data points inside a polygonal region as test statistics. In the latter case, we also show how adding constraints affects the significance of patterns.

3.2.1 Scagnostics as Test Statistics. Scagnostics [32] characterise the visual appearance of an entire scatterplot and model, in a sense, the relationships between the points in the scatterplot. We now consider the scatterplot in Fig. 3, showing the attributes `Finance Workforce` and `Population Density` in the `German` dataset.

Suppose that we believe these attributes to be independent (our null hypothesis) and wish to investigate how this assumption is reflected in the scagnostics. After deciding to test for scagnostics, we plot the data (Fig. 3). We observe that the plot appears *skewed* and *monotonic* and compute these scagnostics (column T in Tab. 2). Are these values significant when compared to our assumption that the attributes are independent? Can we determine if these are random artefacts?

To determine the significance of the scagnostics we follow the steps in Sec. 2.2 for one iteration ($t = 1$) using the 9 scagnostics measures as test statistics. Our null distribution corresponds to uniformly sampling R datasets from a distribution over all datasets having the same marginal distributions as the original dataset, with the requirement that all attributes are independent. This particular distribution can be easily realised using the following constrained randomisation approach: starting with the original dataset we permute each attribute independently; we use the randomisation scheme described in [11, 18].

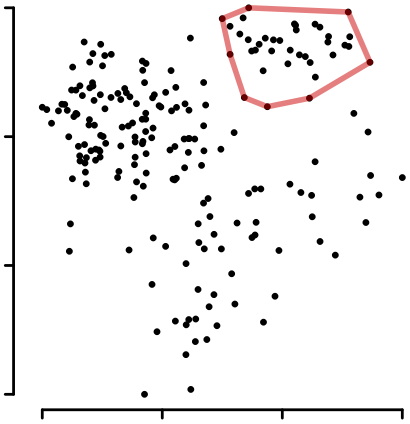
Tab. 2 shows the minP-adjusted p-values for each scagnostic (p_u). Based on the observed p-values we reject the null hypothesis that the attributes in the data are independent for the scagnostic *monotonic*, since its p-value is $\leq \alpha = 0.10$. In other words, if the attributes `Finance Workforce` and `Population Density` are independent, it is unlikely that we would observe this value or higher for the *monotonic* measure. In contrast, for the *skewed* measure we fail to reject the null hypothesis of independence of `Finance Workforce` and `Population Density`.

3.2.2 Effect of Constraints. To demonstrate the effect that constraints have on the p-values, we also compute minP-adjusted p-values for the scagnostics for the scatterplot in Fig. 3 using a null distribution in which the relation between the attributes `Finance Workforce` and `Population Density` is preserved in the R surrogates. These p-values are shown as p_c in Tab. 2. In this case it holds that $T_i(\omega_0) = T_i(\omega_r)$ $(r \in [R])$ for all $i \in [9]$ scagnostics, i.e., the test statistics in the surrogates agree with the original data in this scatterplot, and the p-values cannot be significant.

Next, we consider the scatterplot shown in Fig. 4a, showing a projection of a subset of the `German` data onto the first two principal components. Again we want to examine the null hypothesis that the attributes in the data are independent in terms of the scagnostic

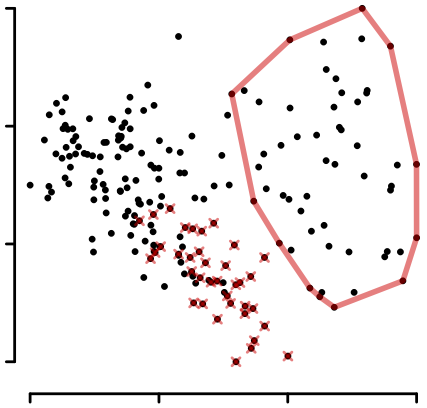
Table 2: Significance of scagnostics computed for the scatterplots in Fig. 3 and Fig. 4a. The columns show the value of the test statistic (T) and the corresponding minP-adjusted p-values for unconstrained (p_u) and constrained null models (p_c). Significant p-values marked with blue.

	Fig. 3			Fig. 4a	
Scagnostic	T	p_u	p_c	T	p_u
Outlying	0.37	0.90	1.00	0.25	0.63
Skewed	0.81	0.95	1.00	0.75	0.52
Clumpy	0.03	0.74	1.00	0.04	0.67
Sparse	0.05	0.80	1.00	0.08	0.03
Striated	0.05	0.95	1.00	0.06	0.94
Convex	0.46	0.91	1.00	0.35	1.00
Skinny	0.46	0.95	1.00	0.55	0.08
Stringy	0.35	0.95	1.00	0.40	0.79
Monotonic	0.63	0.01	1.00	0.01	0.79



	x	w_x	y	w_y
1	Unemploy.	0.30	GREEN.2009	-0.29
2	Unempl..Youth	0.29	Finance.workf.	-0.29
3	Children.Pop.	-0.28	Constr..workf.	0.28
4	LEFT.2009	0.28	Mid.aged.Pop.	-0.27
5	FDP.2009	-0.25	Old.Pop.	0.27

(a)



	x	w_x	y	w_y
1	Children.Pop.	-0.28	Elderly.pop.	-0.40
2	X.cars	-0.27	Unemploy.	-0.35
3	Service.workf.	0.25	No.school.degree	-0.31
4	Prod..workf.	-0.24	Mid.aged.Pop.	0.30
5	LEFT.2009	0.24	SPD.2009	-0.29

(b)

Figure 4: Projected views of `German` data. The tables below the scatterplots provide the attributes with the five largest (in absolute value) weights on the projection axes. (a) In this projection onto the first two principal components we see a significant pattern (marked by a polygonal region shown using solid red lines). (b) Projection of the data onto the *most informative view* as defined in [18] wrt. the pattern in (a) as a constraint on the background distribution. The items in the polygon in (a) are here marked with red crosses. Another significant pattern is marked by a polygonal region with solid red lines.

measures. The values of the test statistics (T) and the corresponding minP-adjusted p-values (p_u) are shown in Tab. 2. The p-values indicate that the null hypothesis concerning the independence of the attributes can be rejected in the case of the *sparse* and *skinny* scagnostics. In other words, if the axes of Fig. 4a are independent, then it is unlikely that we would observe these values or higher for the *sparse* and *skinny* scagnostics.

3.2.3 Iterative Exploration. Instead of scagnostics, which consider the global structure of the entire scatterplot, we now turn to a different test statistic which can be used to investigate local structures. We also consider iterative exploration, which requires MCC between iterations, i.e., the p-values are adjusted according to Eq. (1). The data used here is splittable, i.e., the first part is used for exploration and formulating hypotheses and the second part for testing the hypotheses. Now, the projection of the `German` data onto the first two principal components shown in Fig. 4a seems to contain interesting local cluster patterns. One pattern, for instance, is shown with a solid red polygonal line in Fig. 4a and represents a selection of *rural* districts located in the *East*, which have, e.g., high values for attributes related to unemployment and low values for the attribute describing the voting for the Green party in the 2009 elections (more details provided in the table below Fig. 4a).

We now want to determine if the cluster visible in this polygon marked by the analyst is a true representation of the relations between the attributes in the data or whether it is just a spurious artefact. As our null distribution we again use samples where the attributes of the original dataset have been permuted independently. In this first exploratory step we find that the pattern inside the polygon is indeed significant since the p-value corrected for the iteration ($t = 1$) is 0.002. This practically means that the observed cluster is not present in datasets which have been sampled from the null distribution (which encodes our assumptions about the data) and hence is unlikely to be a random artefact.

We continue the experiment by making a further exploration step. Having observed the significant pattern in the polygon in Fig. 4a, we now wish to continue the exploration of the data and we hence add the observed pattern as a *constraint* to our *background distribution*. We follow the methodology for constrained randomisation of tabular data presented in [11, 18], where tabular data is permuted using *tile constraints*. Tile constraints allow us to generate surrogates for tabular data such that for a subset of items (here: the points inside the polygon) the interaction between certain attributes (here: all attributes in the dataset) is retained. Using this method, the observed pattern in the polygon is constrained (fixed) in the surrogates, whereas all points outside the polygon can be permuted independently. Using this null distribution the p-value of the observed pattern in Fig. 4a equals unity (i.e., non-significant), as expected, since our constraint explains the observed pattern.

Having incorporated the observed pattern into our background distribution we determine a new projection using the method presented in [18], allowing us to compute the so-called *most informative view* of the German data with respect to the background distribution, shown in Fig. 4b. The items originally inside the polygon in Fig. 4a are marked with red crosses for illustration purposes, and it is clear that these points no longer present an interesting pattern in the new view. We observe a pattern consisting of the points in the right side of the projection (selection inside the red polygon). These points are *urban* districts, which have, e.g., lower values in attributes related to the amount of children and the number of cars, and higher values in the attributes describing the percentage of service workforce and voting for the Left party in the 2009 elections. We compute the iteration-adjusted p-value using Eq. (1) with $t = 2$ for this pattern and obtain 0.004, i.e., also this pattern is significant. We add a tile constraint for this pattern, and update the null distribution, after which the pattern is non-significant (p-value 1.00).

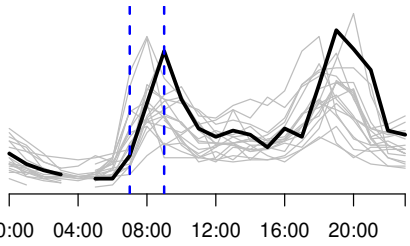


Figure 5: Time series showing the hourly level of CO in the `Air Quality` dataset on March 17th, 2004. The x-axis shows the time of the day. The gray lines are a random sample of other days (historical surrogates). An interesting interval is observed between times 7:00 and 9:00 in the morning, where there appears to be an unusually high increase in the level of CO. The hourly average value data point for 4:00 is missing.

3.3 Testing Hypotheses in Time Series

We already demonstrated how to determine the significance of patterns for time series in *Motivating Example 1* in the introduction, using typical and constrained surrogates obtained using a Gaussian process. We here provide a second example using the `Air Quality` dataset with a different test statistic and *historical surrogates*.

Fig. 5 shows the level of carbon monoxide (CO) on Tuesday, March 17th, 2004. We observe an unusually large increase in the level of CO during 07:00–09:00 in the morning which could be due to, e.g., increased traffic during the morning rush. We want to investigate if the observed increase is a true phenomenon in the `Air Quality` data, or is it just a random observation found only in the particular time series for March 17th, 2004 that we explore? Our null hypothesis is hence that this increase in the level of CO between 07:00 and 09:00 is typical in the data. To test this hypothesis, we use the other business days (i.e., exclude Saturdays and Sundays) as surrogates and the difference between the level of CO at 09:00 and 07:00 as a test statistic. We then compute the values of the test statistic for all two-hour intervals to obtain the minP-adjusted p-values. We find that the unadjusted p-value of the pattern we visually observed in the time series is 0.03 while the adjusted value is 0.67. We hence conclude that we do not have enough evidence to reject our null hypothesis, i.e., the increase in the level of CO between 07:00 and 09:00 is not unusually high in the morning of March 17th, 2004. Observe that the unadjusted p-value of the pattern (i.e. the "baseline") lead to the opposite conclusion.

3.4 Scalability

One possible use case for the significance testing framework presented in this paper is interactive visual exploratory data analysis. This means, that the analyst exploring data should be able to test hypotheses during the data exploration. For this to be possible, it is essential that the significance testing procedure is fast enough. The most time consuming operation in the above framework is the generation of surrogate data. All experiments in this paper can be run in less than 10 minutes using a standard Apple MacBook Pro with a 3.1 GHz Intel Core i5. The analysis of, e.g., the time series example in Fig. 1 takes on the order of a few seconds, which clearly is fast enough for interactive use. As a rough estimate, the time to test a single pattern is $R \times (T_T + T_S) + T_C$ where R is the number of surrogates, T_T and T_S are respectively the computation time for the test statistic T and sampling a surrogate dataset S, and T_C is the time for applying MCCs (minP and iteration adjustment).

It should be noted that the generation of surrogates totally depends on the null hypothesis, and for certain complex hypotheses it is possible that, e.g., Markov Chain Monte Carlo (MCMC) methods must be used to generate surrogates (e.g., [10]), which can be computationally demanding.

4 DISCUSSION

In this paper we have presented a principled framework for evaluating the significance of visual patterns during exploration. The significance of a visual pattern can be thought of as how likely it is to have occurred by chance if we assume a certain distribution for the data (null distribution). The pattern is quantified through a test statistic (e.g., the number of points inside a region) and the null distribution models the user's assumptions about the data (e.g., the independence of some attributes).

We empirically demonstrated how the statistical significance of patterns is influenced by the knowledge of the user (Sec. 3.1). Furthermore, we showed that the framework can be used in the analysis of different types of data using both tabular and time series data. We used different types of null distributions (typical and constrained realisations as well as historical surrogates) and different test statistics. Depending on the hypothesis being tested and on the type of surrogate data being used, we demonstrated how multiple comparisons corrections must be used. Our framework is hence applicable in many different data analysis scenarios and represents an important contribution in exploratory data analysis by making it possible to directly determine the significance of visually observed patterns. Furthermore, we have shown that the statistical significance can be evaluated during iterative data exploration. Exploratory data analysis is a highly important process, the success of which has an important impact on further analyses and modelling of the data, e.g., using machine learning algorithms. Our framework bridges the gap between exploratory and confirmatory analysis by

making it possible to evaluate the significance of patterns found visually during exploration.

When considering patterns in a dataset an analyst considers the absolute values of the observed patterns. However, it is also important to consider both the practical and statistical significance of the patterns, as exemplified in the experiments above. The proposed framework makes this possible. The framework is clearly fast enough for interactive use. In the future we hope to perform comprehensive user studies, in which we study how users investigate visual patterns in the data and how the proposed framework could be best integrated into data analysis workflows. In this paper, we have discussed rather generic problems, and an interesting avenue for future research is to investigate how the ideas presented here could be implemented concretely in some specific domain, e.g., in the analysis of networks or time series. Each domain involves different choices of test statistics and visual representations.

Summarising, the framework proposed in this paper represents a novel and principled method for determining the significance of visual patterns, which can be applied in a vast number of exploratory data analysis scenarios.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (decisions 326280 and 326339).

REFERENCES

[1] Anushka Anand, Leland Wilkinson, and Tuan Nhon Dang. 2012. Visual Pattern Discovery using Random Projections. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 43–52. https://doi.org/10.1109/vast.2012.6400490

[2] Mario Boley, Michael Mampaey, Bo Kang, Pavel Tokmakov, and Stefan Wrobel. 2013. One Click Mining—Interactive Local Pattern Discovery through Implicit Preference and Performance Learning. In *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA)*. ACM, 27–35. https://doi.org/10.1145/2501511.2501517

[3] Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, 1906 (2009), 4361–4383. https://doi.org/10.1098/rsta.2009.0120

[4] Tuan Nhon Dang, Anushka Anand, and Leland Wilkinson. 2013. TimeSeer: Scagnostics for High-Dimensional Time Series. *IEEE Transactions on Visualization and Computer Graphics* 19, 3 (2013), 470–483. https://doi.org/10.1109/tvcg.2012.128

[5] Michael de Ridder, Karsten Klein, and Jinman Kim. 2018. A Review and Outlook on Visual Analytics for Uncertainties in Functional Magnetic Resonance Imaging. *Brain Informatics* 5, 2 (2018), 5. https://doi.org/10.1186/s40708-018-0083-0

[6] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. 2008. On Field Calibration of an Electronic Nose for Benzene Estimation in an Urban Pollution Monitoring Scenario. *Sensors and Actuators B: Chemical* 129, 2 (2008), 750–757. https://doi.org/10.1016/j.snb.2007.09.060

[7] Dheeru Dua and Casey Graff. 2019. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[8] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. 2003. Multiple Hypothesis Testing in Microarray Experiments. *Statist. Sci.* 18, 1 (2003), 71–103. https://doi.org/10.1214/ss/1056397487

[9] Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. 2006. False Discovery Control with p-Value Weighting. *Biometrika* 93, 3 (2006), 509–524. http://www.jstor.org/stable/20441304

[10] Andreas Henelius, Jussi Korpela, and Kai Puolamäki. 2013. Explaining Interval Sequences by Randomization. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. LNCS, Vol. 8188. Springer, 337–352. https://doi.org/10.1007/978-3-642-40988-2_22

[11] Andreas Henelius, Emilia Oikarinen, and Kai Puolamäki. 2019. Tiler: Software for Human-Guided Data Exploration. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. LNCS, Vol. 11053. Springer, 672–676. https://doi.org/10.1007/978-3-030-10997-4_49

[12] Bo Kang, Jefrey Lijffijt, Raúl Santos-Rodríguez, and Tijl De Bie. 2016. Subjectively Interesting Component Analysis: Data Projections that Contrast with Prior Expectations. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1615–1624. https://doi.org/10.1145/2939672.2939840

[13] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2009. Visual Analytics: Scope and Challenges. In *Visual Data Mining Theory, Techniques and Tools for Visual Analytics*. Springer, 76–90. https://doi.org/10.1007/978-3-540-71080-6_6

[14] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI, 202–207.

[15] Jefrey Lijffijt, Panagiotis Papapetrou, and Kai Puolamäki. 2014. A Statistical Significance Testing Approach to Mining the Most Informative Set of Patterns. *Data Mining and Knowledge Discovery* 28, 1 (2014), 238–263. https://doi.org/10.1007/s10618-012-0298-2

[16] Mahbubul Majumder, Heike Hofmann, and Dianne Cook. 2013. Validation of Visual Statistical Inference, Applied to Linear Models. *J. Amer. Statist. Assoc.* 108, 503 (2013), 942–956. https://doi.org/10.1080/01621459.2013.808157

[17] Kristin Potter, Joe Kniss, Richard Riesenfeld, and Christopher R. Johnson. 2010. Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum* 29, 3 (2010), 823–832. https://doi.org/10.1111/j.1467-8659.2009.01677.x

[18] Kai Puolamäki, Emilia Oikarinen, and Andreas Henelius. 2019. Guided Visual Exploration of Relations in Data Sets. *CoRR* abs/1905.02515 (2019), 30 pages. https://arxiv.org/abs/1905.02515

[19] Kai Puolamäki, Panagiotis Papapetrou, and Jefrey Lijffijt. 2010. Visually Controllable Data Mining Methods. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*. IEEE, 409–417. https://doi.org/10.1109/ICDMW.2010.141

[20] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[21] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 240–249. https://doi.org/10.1109/tvcg.2015.2467591

[22] Thomas Schreiber and Andreas Schmitz. 2000. Surrogate Time Series. *Physica D: Nonlinear Phenomena* 142, 3 (2000), 346–382. https://doi.org/10.1016/S0167-2789(00)00043-9

[23] Lin Shao, Timo Schleicher, Michael Behrisch, Tobias Schreck, Ivan Sipiran, and Daniel A. Keim. 2016. Guiding the Exploration of Scatter Plot Data using Motif-based Interest Measures. *Journal of Visual Languages & Computing* 36 (2016), 1–12. https://doi.org/10.1016/j.jvlc.2016.07.003

[24] Julian Stahnke, Marian Dörk, Boris Müller, and Andreas Thom. 2016. Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 629–638. https://doi.org/10.1109/tvcg.2015.2467717

[25] James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. 1992. Testing for Nonlinearity in Time Series: The Method of Surrogate Data. *Physica D: Nonlinear Phenomena* 58, 1–4 (1992), 77–94. https://doi.org/10.1016/0167-2789(92)90102-S

[26] James Theiler and Dean Prichard. 1996. Constrained-Realization Monte-Carlo Method for Hypothesis Testing. *Physica D: Nonlinear Phenomena* 94, 4 (1996), 221–235. https://doi.org/10.1016/0167-2789(96)00050-4

[27] Niko Vuokko and Petteri Kaski. 2011. Significance of Patterns in Time Series Collections. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM)*. SIAM, 676–686. https://doi.org/10.1137/1.9781611972818.58

[28] Peter H. Westfall and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley.

[29] Hadley Wickham, Dianne Cook, and Heike Hofmann. 2015. Visualizing Statistical Models: Removing the Blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8, 4 (2015), 203–225. https://doi.org/10.1002/sam.11271

[30] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. Graphical Inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 973–979. https://doi.org/10.1109/tvcg.2010.161

[31] Holly M. Widen, James B. Elsner, Stephanie Pau, and Christopher K. Uejio. 2015. Graphical Inference in Geographical Research. *Geographical Analysis* 48, 2 (2015), 115–131. https://doi.org/10.1111/gean.12085

[32] Leland Wilkinson, Anushka Anand, and Robert Grossman. 2005. Graph-Theoretic Scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS)*. IEEE, 157–164. https://doi.org/10.1109/INFOVIS.2005.14

[33] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 479:1–479:12. https://doi.org/10.1145/3173574.3174053