Uni- and Multivariate Probability Density Models for Numeric Subgroup Discovery

Marvin MEENG [a], Harm DE VRIES [b], Peter FLACH [c], Siegfried NIJSSEN [d] and Arno KNOBBE [a]

[a] *LIACS, Leiden University, the Netherlands* {meeng,knobbe}@liacs.nl
[b] *Université de Montréal, Canada* mail@harmdevries.com
[c] *Bristol University, United Kingdom* Peter.Flach@bristol.ac.uk
[d] *Université Catholique de Louvain, Belgium* siegfried.nijssen@uclouvain.be

Abstract. Subgroup Discovery is a supervised, exploratory data mining paradigm that aims to identify subsets of a dataset that show interesting behaviour with respect to some designated target attribute. The way in which such distributional differences are quantified varies with the target attribute type. This work concerns continuous targets, which are important in many practical applications. For such targets, differences are often quantified using z-score and similar measures that compare simple statistics such as the mean and variance of the subset and the data. However, most distributions are not fully determined by their mean and variance alone. As a result, measures of distributional difference solely based on such simple statistics will miss potentially interesting subgroups. This work proposes methods to recognise distributional differences in a much broader sense. To this end, density estimation is performed using histogram and kernel density estimation techniques. In the spirit of Exceptional Model Mining, the proposed methods are extended to deal with *multiple* continuous target attributes, such that comparisons are not restricted to univariate distributions, but are available for joint distributions of any dimensionality. The methods can be incorporated easily into existing Subgroup Discovery frameworks, so no new frameworks are developed.

Keywords. Subgroup Discovery, Continuous Target, Complex Target Concepts, Kernel Density Estimation

1. Introduction

Subgroup Discovery is a supervised, exploratory data mining paradigm that aims to identify subsets of a dataset that show interesting behaviour with respect to some designated target attribute [1–4]. Although exploration strategies and optimisations may differ, all Subgroup Discovery algorithms have in common that they generate candidate subgroups and assign them a score using a quality measure. These subgroups are formed by interpretable descriptions that select subsets of the data through simple constraints on that data. Quality measures typically score subgroups by quantifying the difference between the distribution on the target for subsets of the data and the overall dataset. A wide array of quality measures is available, and these are categorised by type of the target attribute.

Although the nominal target type is the most common, many data sources sport a numeric type. In the numeric target setting, in order to compare (continuous) distributions

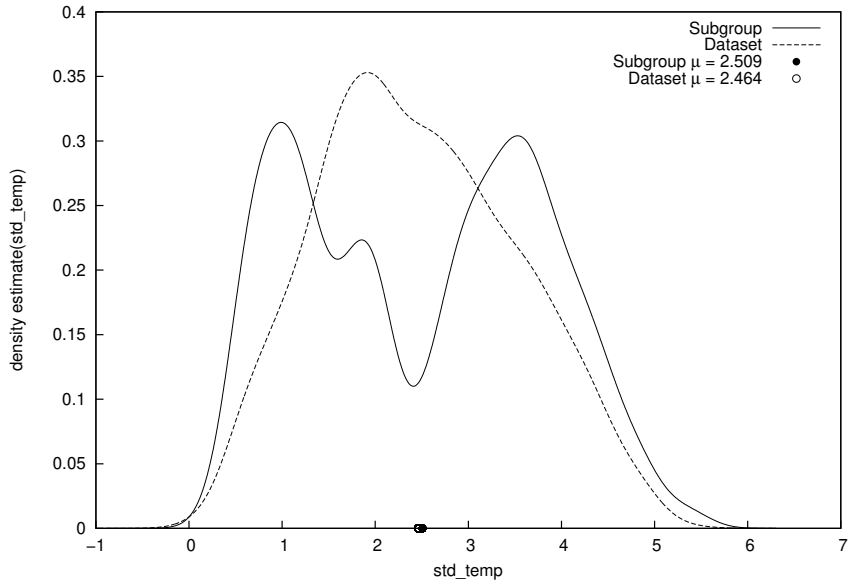


Figure 1. Real-world example of two very distinct distributions with almost identical means, indicated by the points on the x-axis. Quality measures based on a simple statistic such as the mean would not judge the two distributions as very dissimilar.

over target attributes, one first needs to estimate these distributions. But, in contrast to the discrete case, there are many potential target values that will never be observed. This makes the discovery of subgroups challenging, the main difficulty being that only a limited number of point measurements of these distributions is available.

Some earlier works have studied the problem of finding subgroups in data with a numeric target attribute. Typical solutions look for subgroups where the average value of the target attribute is surprisingly high (or low) compared to that of the overall dataset. Quality measures judge a subgroup based on simple statistics such as mean and variance, typically getting their inspiration from well-known statistical tests. Popular examples include the measures *mean*-test, z-score and t-statistic [5–7]. In this approach, the point measurements are hence used to estimate the parameters of a simple distribution (such as a normal distribution). However, the assumptions that are made in such an approach with respect to the underlying distribution may not always be valid.

As an example, consider Figure 1 for an illustration of a subgroup in real-world data that is not recognised by traditional measures. In this example, the subgroup has the same mean as the overall dataset, but the target distribution in the subgroup is bimodal whereas it is unimodal in the overall dataset. The first challenge addressed in this work is how to score the quality of subgroups such that subgroups of this kind can be identified.

The second challenge is the discovery of subgroups in data with multiple numeric targets, a setting commonly referred to as Exceptional Model Mining [8]. Specifically, the interest is in capturing subgroups that relate to unusual differences in the *joint* distribution over the multiple target attributes. Therefore, simply treating the target attributes independently is not an option.

To address these challenges, measures are proposed that can quantify subtle differences in distribution, even when multiple target attributes are involved:

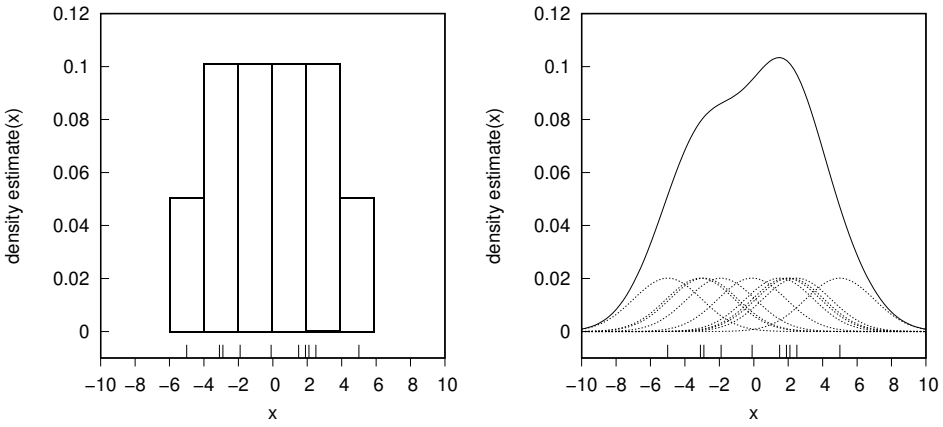


Figure 2. Example of density estimation using a histogram (left) and KDE (right), data points are indicated by a tic on the bottom axis. For the histogram, when the value of an observation falls between the bin boundaries of a bin, the relative frequency for this bin is increased. For the KDE technique, the final estimate is obtained by placing a kernel on each observation, and then summing the result.

- the first class of measures is based on density estimation using histograms; and
- the second class of measures is based on Kernel Density Estimation (KDE), also known as the Parzen-Rosenblatt window method [9, 10] in some fields.

Figure 2 illustrates the differences between these techniques. In the histogram technique, the range of values of a target attribute is discretised in bins, after which the relative frequency for each bin is determined, yielding density estimate histograms that are then compared. For the KDE technique, a smoothing kernel is placed over each observation and the resulting smooth representations of distributions are compared. In this context, two distance measures between distributions are introduced. The first uses numerical approximation to compute the (approximate) probability density functions, and serves as a general example of methods for which no analytical solution is known. For the other distance an analytical solution exists, offering a much reduced computational complexity, as one can avoid the creation of distributions altogether.

KDE-based techniques have the potential to address some of the weaknesses of histogram-based techniques: the unfortunate choice of bin boundaries and the number of bins [11]. Most importantly, they do not require the careful tuning of a discretisation algorithm, instead requiring the definition of a kernel.

So, the proposed solution compares complete distributions, instead of just simple statistics. Limited work exist that utilises quality measures gauging the difference in distribution in a broader sense [6, 12]. Both aforementioned works use the Kolmogorov-Smirnov test, which returns the maximum distance between the (empirical) distribution of the subset, and the (cumulative) distribution of the overall dataset. The Kolmogorov-Smirnov test compares two cumulative distribution functions, and could therefore be considered more complex than a simple statistic. However, it still only considers, and returns, just a single aspect of these distributions, the largest difference between them, occurring at any (single) point of the distributions. Furthermore, it is not straightforward to extend the Kolmogorov-Smirnov test to multiple dimensions, although some variants have been proposed [13, 14]. The aforementioned Subgroup Discovery works using the

Kolmogorov-Smirnov test did not extend their solution to multiple dimensions, whereas the solution proposed in this work does. Therefore, it is not only broader in the sense that it compares complete distributions, instead of simple statistics. It also applies to, what are called, traditional Subgroup Discovery and Exceptional Model Mining.

In summary, the contributions of this work are as follows: several measures are introduced to evaluate the quality of a subgroup in data with possibly multiple numeric target attributes, based on histograms and kernel density estimation, and, for these measures, parameter tuning, and how they compare to each other in practice, is evaluated.

Section 2 addresses related work, followed by Section 3, which introduces the required notation, and discusses earlier methods. Then follows a discussion of histogram-based measures in Section 4 and KDE-based measures in Section 5. The measures are experimentally evaluated in Section 6, the implications of the experiments are discussed in Section 7, and Section 8 presents a conclusion.

2. Related Work

This section discusses related work in the domains of Subgroup Discovery and Exceptional Model Mining. Literature with respect to discretisation, density estimation and bandwidth selection is out of the scope of this discussion, though relevant work is referenced throughout the text.

A number of researchers have proposed methods to deal with numeric target attributes in Subgroup Discovery. The majority of work gauges interestingness with respect to the target using simple statistics. For example, [15] introduces a mean-based solution, that involves the expensive process of copying the dataset and running the Subgroup Discovery algorithm multiple times. The current work goes beyond simple statistics, as it was shown that interesting subgroups will be overlooked when relying on them.

Other work [6,16,17] has focussed more on efficient search space exploration. However, efficiency gains often come from global discretisation [18] of numeric discription attributes [5, 6, 16, 19]. This pre-discretisation essentially turns a numeric feature into a nominal one, leading to a much reduced search space for which bounds can be computed and stored. Though not essential to the new density estimation model class, the current work employs a richer description language that allows for any interval to be created *during* search space exploration, as to best adapt to the target domain, a feature that has been shown to increase performance for other model classes [20], but cannot efficiently be integrated in the algorithm of the aforementioned works. Still, the interesting work of [16] extends the optimistic estimator framework [4–6, 15, 17, 21, 22] to include efficiently computed tight estimators for all functions that are determined by the size, median and dispersion around the median of a subgroup. As such, the solution deals with the parameters size, central tendency and dispersion, statistics that do not fully determine most distributions. In this respect, [6] is especially interesting, as it goes beyond simple statistics. Like in [12], Kolmogorov-Smirnov is proposed as measure to quantify the interestingness of the target distribution. Kolmogorov-Smirnov could be considered more than just a simple statistic, as it compares two cumulative distribution functions (CDFs). Still, it is not all that sophisticated, as it returns just the maximum difference between two CDFs, which is thus a single point difference measurement. Clearly, the density difference approach based on comparison of complete distributions presented in the current work goes beyond this limited measurement.

Moreover, a major difference between the current work and those mentioned above, is the extension to the Exceptional Model Mining framework [8]. Several model classes have been proposed using multiple numeric target attributes, including correlation [8] and simple linear [8] and more complex [23] regression models. All of these would be ill-suited to deal with multimodal distributions such as that of the subgroup presented in Figure 1.

Another work that should be noted in the context of Exceptional Model Mining is [22]. It introduces GP-trees, an adaptation of FP-trees [24], that allow for succinct representations of subsets using so-called *valuation bases*, that can be used to optimise search space exploration (though this again requires pre-discretisation). A valuation basis is defined as: 'a (condensed) representation of a set of data instances that is sufficient to extract the model parameters for a given model class.' In connection to the current work, construction of a condensed valuation basis is possible for the histogram-based density estimations techniques. Such a basis would store the histogram bounds or bins, and (relative) counts. Unfortunately, construction of a condensed valuation basis is not possible for the KDE variant. The authors note that: 'If the constructed model itself is very complex, then it seems difficult to derive suitable condensed valuation bases that are sufficient to extract the model parameters', and they give the learning of a bayesian network [25] as an example. To represent the distribution or density estimate for a subgroup, storing just the kernel and bandwidth h is insufficient, and the density estimate can only be created by including all individual instances, leading to, what the authors describe as, a trivial valuation basis. An example is given of a single target, variance valuation basis, which could serve as (input for) bandwidth h, but besides the fact that multidimensional (co)variance matrices would be required, the step computing the density estimate would violate the single-pass constraint imposed by the algorithm. Finally, to avoid confusion, the work includes an example of a valuation basis for DTM-classifiers, that uses the Hellinger distance, but it is in connection to a discrete target, so of no concern to the work presented here.

3. Preliminaries

Throughout this work, the following definition of a dataset is used:

Definition 1. (Dataset) A dataset $\mathscr{D}$ is a bag of N *records* $\vec{r}^i \in \mathscr{D}$ of the form: $\vec{r}^i = \left(a_1^i, \ldots, a_k^i, t_1^i, \ldots, t_m^i,\right)$, with k and m positive integers from $\mathbb{Z}^+$.

Here, $a_1^i, \ldots, a_k^i$ are the values of $\vec{r}^i$ for the *description attributes* $a_1, \ldots, a_k$, and $t_1^i, \ldots, t_m^i$ are the values for the *target attributes* $t_1, \ldots, t_m$. Description attributes are taken from an unrestricted domain $\mathbb{A}$, but only continuous numeric attributes, taken from $\mathbb{R}$, can serve as target. Vector notation $\vec{(\cdot)}$ is used when collectively referring to a number of related values.

Before moving to the definition of a subgroup, note that a distinction is made between the intensional and extensional part of a subgroup. Informally, the intensional part of a subgroup is its description, and the extension consists of the actual records that make up the subgroup. Descriptions are important in Subgroup Discovery, as it was developed as a descriptive and exploratory data mining paradigm, allowing for the discovery of new and *interpretable* knowledge about data. That a result set might contain multiple differ-

ent descriptions for the exact same set of records is not problematic per se. In fact, this is essential to be able to establish correlations, and, individually, the descriptions might all represent previously unknown concepts and therefore yield new knowledge [26, 27].

Definition 2. (Description) A description is a function: $I : \mathbb{A}^k \to \{0,1\}$, where I *covers* a record $\vec{r}^i$ *iff*:

$$I\left(a_1^i, \ldots, a_k^i\right) = 1.$$

Typically, the *description language* $\mathscr{I}$ in Subgroup Discovery consists of (conjunctions of) conditions on description attributes of the general form 'a_i *operator value*'. Examples include '$Smokes = false$' and '$EyeColour = brown \wedge Length \geq 1.76$'. The *depth* of a subgroup is defined as the number of conjuncts in the description. Often, the maximum depth of a description is constrained by a parameter of the Subgroup Discovery algorithm called search depth, designated by d.

In a sense, one could say a subgroup description *precedes* a subgroup extension, in that it is through the description that a subset of records is selected from the dataset. Definition 3 expresses this relation.

Definition 3. (Extension) An extension $\mathscr{E}_I$ corresponding to a description I is the bag of records $\mathscr{E}_I \subseteq \mathscr{D}$ that I covers:

$$\mathscr{E}_I = \left\{\vec{r}^i \in \mathscr{D} \;\middle|\; I\left(a_1^i, \ldots, a_k^i\right) = 1\right\}.$$

From now on the subscript I is omitted if no confusion can arise, and a subgroup extension is simply referred to as $\mathscr{E}$.

The explicit differentiation of the intensional and extensional facets of a subgroup is required in some Subgroup Discovery algorithms [28]. However, for the remainder of this work, s denotes a subgroup, encompassing both its intension and extension. For any particular subgroup s, with extension $\mathscr{E}$, n denotes its size, that is, the number of records in that subgroup: $n = |\mathscr{E}|$. The complement of a subgroup is denoted by $\mathscr{E}^c$, and its size is denoted n^c. Hence, $\mathscr{E}^c = \mathscr{D} \backslash \mathscr{E}$, and $n^c = N - n$.

A *quality measure* objectively evaluates a candidate description in a given dataset. For each description I in the description language $\mathscr{I}$, a quality measure is a function that quantifies how interesting the subgroup s is.

Definition 4. (Quality Measure) A *quality measure* is a function $\varphi_{\mathscr{D}} : \mathscr{I} \to \mathbb{R}$ that assigns a unique numeric value to a description I and its associated extension $\mathscr{E}_I$, given a dataset $\mathscr{D}$.

In principle, Subgroup Discovery algorithms aim to discover subgroups that score high on a quality measure. However, it is common practice to also impose additional *constraints* on subgroups that are found by Subgroup Discovery algorithms. Usually these constraints include the aforementioned upper bound on the complexity of the descriptions ($d \leq ub$), and lower bounds on the quality of the description ($\varphi_{\mathscr{D}}(I) \geq lb_1$) and the size of the induced subgroup ($n \geq lb_2$). More constraints may be imposed as the question at hand requires.

A number of Subgroup Discovery software packages offer facilities and algorithms for dealing with numeric target attributes [19, 29]. Typically, the quality measures used to

assess the interestingness of a subgroup are based on some simple statistics. For example, the z-score measure quantifies interestingness using the difference between the subgroup and data mean [7]:

$$z\text{-score}_{\mathscr{D}}(I) = \frac{\sqrt{n} \cdot (\mu_s - \mu_{\mathscr{D}})}{\sigma_{\mathscr{D}}}, \tag{1}$$

where I is the subgroup (description) and $\mathscr{D}$ the dataset, n is the subgroup size, μ_s is the mean of the target attribute values in the subgroup, and $\mu_{\mathscr{D}}$ and $\sigma_{\mathscr{D}}$ are the mean and standard deviation of the whole dataset, respectively.

Many different search strategies can be used to identify high-scoring subgroups. The most common search strategies are those that recursively add conditions to descriptions, refining one or more of the best-scoring subgroups until a constraint is violated or no improvement is found. These top-down search strategies differ in the order in which refinements are considered and which descriptions are considered for refinement.

4. Probability Density Estimation and Divergence Using Histograms

The main challenge addressed in this work is to develop principled methods for Subgroup Discovery with one or more continuous targets. This requires representing the empirical distribution of the target(s) over a set of records, such that one can look for surprising subgroups by comparing distributions within the subgroup with the distribution in the overall dataset.

An obvious first approach is to discretise the continuous target(s) by means of *histograms* [30]. This involves splitting the target range $[l,u]$ into B bins requiring $B-1$ bin boundaries. The number of data points within each bin gives the height of the corresponding histogram bar, so that, collectively, the height of the bars adds up to the total number of data points. A normalised histogram divides the height of each bar by the total number of data points, so that the bars represent relative frequencies adding up to 1.

A number of fields have developed discretisation algorithms, among these are the database field [31, 32] and the machine learning field. They have been used in the construction of indexes, for query optimisation, for limiting search spaces over hypotheses [33, 34], or to allow the use of algorithms otherwise not equipped to handle continuous data. As this work deals with continuous targets, possibly multiple of them, supervised discretisation methods can not be employed. Instead, two straightforward, unsupervised, binning strategies are evaluated that allow for fast discretisation of (multidimensional) continuous target domains.

In *equal-width* (EW) binning, the bin boundaries are equally spaced. For a bounded scale, the bin width follows directly from the number of bins required.

Alternatively, one can use quantiles, also called *equal-height* (EH), or equal-frequency, bins. Here, the boundaries are positioned such that the number of observations that fall into each bin is – as much as possible – equal. This alternative increases focus in high density regions, and is less sensitive to outliers.

In this work, both binning processes are evaluated. The number of bins is determined using an informed approach discussed in Section 6.5, and used to establish and fix the bin boundaries prior to the mining process. Subsequently, during the search, whenever a relative frequency is required for a subgroup or complement, these same bin boundaries are used, as differing (number of) boundaries would yield incomparable histograms.

4.1. Quality Measure for Probability Density Functions

Irrespective of the discretisation method used, a histogram-based method creates a discrete distribution, and an accompanying probability mass function (PMF), in contrast to the probability density function used in a continuous KDE setting. Conveniently, distance measures exist that can deal with both PMFs and PDFs, by offering two related, but alternative, variations. So, although many distances and divergences are available in the literature, and essentially any measure that characterises the difference between two *functions* can be used as quality measure, this work focusses on the *squared Hellinger* distance measure, as both a discrete and continuous variant of this distance measure are available. This is convenient, as it allows working with essentially the same measure in both the histogram-based discrete setting and the continuous KDE setting. Within the context of the histogram-based methods, the discrete squared Hellinger distance measure is used:

Definition 5. (Squared Hellinger – discrete)

$$H_d^2(P,Q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2. \tag{2}$$

This distance measure is representative of a wide range of distance measures, such as the Earth Movers or Wasserstein distance [35], or Bhattacharyya distance [36].

4.2. Complexity of the General Histogram-based Solution

The main advantage of a histogram approach is its low computational complexity.

For EW binning, the pre-processing phase consists of two passes over the data, and requires $\mathcal{O}(N \log B)$, with N being the number of records in the data, and B the number of bins. The minimum and maximum of the target domain are established first. Then, bin boundaries are computed such that records can be allocated to the appropriate bins.

EH binning has a complexity of $\mathcal{O}(N \log N)$. First, all data is sorted. Then, bin boundaries are determined, and records are allocated to the appropriate bins.

For both alternatives, the 'total' count for each bin, based on the whole dataset, needs to be established only once, before the actual Subgroup Discovery process commences. More important is the complexity of computing the final density estimate for subgroups, as this is typically performed a large number of times.

If a subgroup is compared to its complement, the estimate for both is required. Nonetheless, this only costs $\mathcal{O}(S')$, where $S' = \min(n, n^c)$, that is, the minimum of the subgroup size and the size of its complement. First, a single pass over the smaller of the two subsets establishes its bin counts. Subtracting these from the 'total' counts for the dataset yields the counts for the other subset.

The computational complexity is the same when comparing a subgroup to all the data. If a subgroup constitutes less than half of the data, a single pass over its records gives the required bin counts directly. Else, a pass over its complement is sufficient to compute the subgroup counts, via subtraction of the 'total' counts. This results in a computational complexity of $\mathcal{O}(S')$, with an upper bound for S' of $N/2$, for all situations.

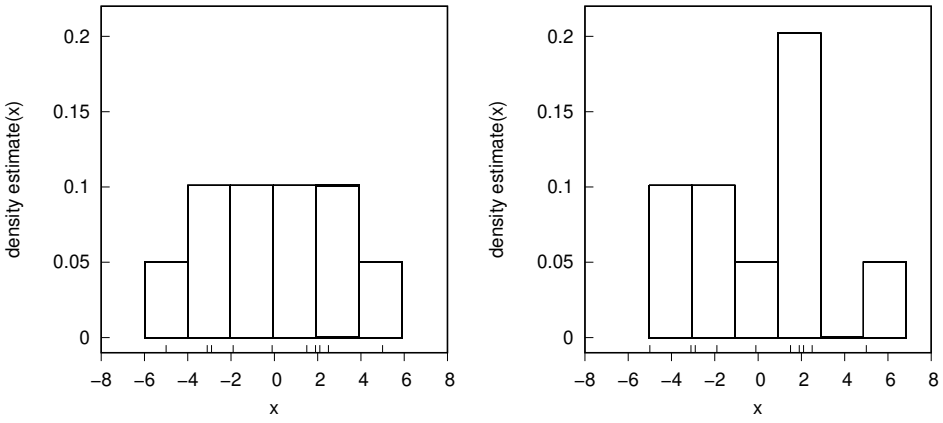


Figure 3. Histogram techniques are sensitive to bin boundary placement. The two histograms above use the same data points, number of bins and bin width, but as a result of the different bin boundary placements, the resulting density estimates are quite distinct.

5. Kernel Density Estimation

As pointed out earlier, the disadvantage of histogram-based techniques is that the results of these techniques can depend very much on the choice of histogram boundaries. The histograms in Figure 3 use the same data points, number of bins and bin width, but as a result of the different bin boundary placements, the resulting density estimates are quite distinct. Kernel Density Estimation addresses this weakness by not requiring the use of bins or bin boundaries. Instead, KDE puts a kernel on each observation and sums the resulting kernels over all observations. The resulting representation of the distribution is smoother and less sensitive to the choice of parameter settings.

The main parameter in this process is the kernel bandwidth parameter h. This parameter essentially indicates how wide the kernel is and hence determines the smoothness of the representation: a larger h will lead to a result that is more smooth. Its continuous nature means that a small change in the parameter setting never leads to a large change in the representation of the distribution. This makes KDE a more robust approach.

More formally, let $K(\cdot)$ denote a symmetric kernel function and h a smoothing parameter, then a kernel density estimator $\hat{f}$ is defined as:

$$\hat{f}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\vec{x}-\vec{x}_i), \tag{3}$$

where $\int K_h(x)\,\mathrm{d}x = 1$ in order to ensure that the probability density estimate sums to 1. In other words, a KDE spreads out the probability mass of each observation by convolving it with a kernel with bandwidth h.

In this general form, two choices still need to be made: the kernel function used ($K(\cdot)$) and the bandwidth parameter used (h). Studies have shown that the actual kernel used is of less influence than the selected bandwidth [37]. A plethora of kernel functions have been proposed which, among others, include Epanechnikov, uniform and triangular [11]. Here, a Gaussian kernel is used, because it is one of the most used kernels, and it allows for an analytic L^2 distance between two KDEs (see Section 5.3). It is stressed,

however, that, in the general concept, any other kernel can be employed, were one so inclined. The Gaussian kernel is equivalent to a multivariate normal distribution, with $\vec{x}$ being a m-dimensional column vector, where m represents the number of target attributes:

Definition 6. (Gaussian kernel)

$$K_h(\vec{x}) = \frac{1}{\sqrt{(2\pi)^m |h|}} \exp\left(-\frac{1}{2}\vec{x}^T h^{-1} \vec{x}\right). \tag{4}$$

Hence, the bandwidth parameter h of this kernel essentially corresponds to the covariance matrix of a normal distribution [38]. h is required to be positive definite, and $|h|$ indicates its determinant. Note that for a single target attribute the target dimensionality is equal to one, so this reduces to a univariate normal distribution.

The main choice to make is which bandwidth to use. As a pragmatic choice, Silverman's rule of thumb is chosen, which is considered to be the most used bandwidth estimator. It is known not to be optimal [37], but its low computational complexity is appreciated. One would like to keep the estimation step relatively light, as a new PDF needs to be constructed for each tested subgroup (and accompanying complement), and the number of tests could run in the millions.

In particular, the variant that takes into account the full covariance matrix Σ of the target attributes is used:

Definition 7. (Silverman's bandwidth estimator)

$$h = \left(\frac{4}{m+2}\right)^{\frac{1}{m+4}} n^{\frac{-1}{m+4}} \Sigma^{\frac{1}{2}}. \tag{5}$$

The estimator transforms the sample covariance matrix Σ into the bandwidth matrix h, used in Eq. 4, by applying some scaling factor. Σ is required to be positive definite (and thus full rank and with non-zero determinant) in order to ensure that the distribution is non-degenerate. This prevents Σ from being singular, which would occur when *all* samples lie in a low-dimensional subspace of the target distribution, for example when they all have the same value for a one-dimension target. For a singular matrix, the determinant is zero and no inverse can be computed[1]. Besides, it would also lead to division by zero in the scale factor of Eq. 4.

5.1. Measuring Density Differences Based on KDE

The kernel density estimates of two distributions can be compared with each other using many well-known distance measures. This work considers the squared Hellinger distance and the L^P distance. More specifically, L^2 is used, because it allows for an analytic solution to the integral as discussed in Section 5.3.

Definition 8. (Squared Hellinger – continuous)

$$H_c^2(P,Q) = \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x \tag{6}$$

[1] In case of a singular covariance matrix, one can use the (Moore-Penrose) pseudoinverse.

Definition 9. (L^2 distance)

$$L^2(P,Q) = \left(\int |p(x) - q(x)|^2 \, \mathrm{d}x \right)^{1/2} \tag{7}$$

This work is restricted to distances, and does not consider divergences, like the well-known Kullback-Leibler divergence [39]. Divergences, in contrast to distances, are not necessarily symmetric, potentially yielding different scores when comparing subgroup versus complement and complement versus subgroup. However, if one considers this not to be a problem, or if symmetric divergences like Jensen-Shannon [40] are used, there is no objection to the inclusion of divergences as a quality measure in the proposed setting.

From a practical point of view, an important issue in the use of the above measures is that they require the calculation of an integral. However, there is, for a general divergence, no analytic solution to this integral. In this work, the integral is approximated numerically, using the rectangle rule and a fixed step-size to set up a (higher-dimensional) grid. The step size δ is based on the minimum of the variance of the subgroup or the variance of the complement, small enough to yield smooth estimates[2]. For each point $\vec{x}$ on the grid, the probability density ($\hat{f}(\vec{x})$), and divergence, are computed. Summing the results of all points then gives the approximation to the quality measures.

5.2. Complexity of General KDE-based Solution

Let G denote the number of grid points (used for the approximation of the integral), and S the size of the subgroup (n), its complement (n^c) or data (N) for which the PDF is computed. Then the computational complexity of the naive method for evaluating the divergence for a one-dimensional PDF is $\mathcal{O}(GS)$. Extending this result to higher dimensions, the number of grid points grows exponentially in the number of dimensions m, yielding $\mathcal{O}(G^m S)$. Indeed, the asymptotic performance of the proposed method is not good for higher dimensions.

5.3. Analytic Solution of L^2 Distance

While in the general case no analytical solution is known and hence numerical integration is necessary, this section will show that in the case of the L^2 distance, one can avoid the naive approximation presented in the previous section. Although the result dates back to [41], its solution is restated because of its importance to speed up Subgroup Discovery for higher-dimensional target distributions. The analysis starts from kernel density estimators of the form:

$$\hat{f}(\vec{x}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(\vec{x}|\vec{v}_i, V) \quad \text{and} \quad \hat{g}(\vec{y}) = \frac{1}{M} \sum_{i=1}^{M} \mathcal{N}(\vec{y}|\vec{w}_i, W), \tag{8}$$

and are interested in their squared L^2 distance:

[2] A δ of $h/8$ yielded smooth enough estimates for the experiments.

$$L^2(\hat{f},\hat{g}) = \int (\hat{f}(x) - \hat{g}(x))^2 \,\mathrm{d}x \tag{9a}$$

$$= \int \hat{f}^2(x)\,\mathrm{d}x - 2\int \hat{f}(x)\hat{g}(x)\,\mathrm{d}x + \int \hat{g}^2(x)\,\mathrm{d}x. \tag{9b}$$

This represent the kernels as Gaussians $\mathcal{N}(\vec{x}|\vec{v}_i, V)$, where $\vec{v}_i$ indicates the mean of the Gaussian and V indicates a covariance matrix, corresponding to the bandwidth parameter of the kernel.

The following only provides a solution for the dot product $\int \hat{f}(x)\hat{g}(x)\,\mathrm{d}x$, however, the derivation can be applied to the L^2 norm of the functions $\hat{f}$ and $\hat{g}$ in a similar way. Consecutively plug in the kernel density estimators, expand the quadratic expression and apply the sum rule of integration to obtain:

$$\int \hat{f}(x)\hat{g}(x)\,\mathrm{d}x = \frac{1}{NM}\sum_{i,j=1}^{N,M} \int \mathcal{N}(\vec{x}|\vec{v}_i, V)\mathcal{N}(\vec{x}|\vec{w}_j, W)\,\mathrm{d}x. \tag{10}$$

The product of two Normal distributions is an (unnormalised) Gaussian [42, A.7]:

$$\mathcal{N}(\vec{x}|\vec{v}, V)\mathcal{N}(\vec{x}|\vec{w}, W) = Z^{-1}\mathcal{N}(\vec{x}|\vec{\upsilon}, \Upsilon), \tag{11a}$$

$$\text{where} \quad \vec{\upsilon} = \Upsilon(V^{-1}\vec{v} + W^{-1}\vec{w}) \tag{11b}$$

$$\text{and} \quad \Upsilon = (V^{-1} + W^{-1})^{-1}. \tag{11c}$$

Here the normalising constant is:

$$Z^{-1} = (2\pi)^{-m/2}|V+W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\vec{v}-\vec{w})(V+W)^{-1}(\vec{v}-\vec{w})^{\top}\right), \tag{12}$$

which is exactly the solution to the integral:

$$\int Z^{-1}\mathcal{N}(\vec{x}|\vec{\upsilon}, \Upsilon)\,\mathrm{d}x = Z^{-1}. \tag{13}$$

Plugging the above result in Eq. 10 gives the solution to the integral of the dot product of $\hat{f}$ and $\hat{g}$. As mentioned earlier, it is straightforward to derive the L^2 norm of $\hat{f}$ and $\hat{g}$ (the first and third term in Eq. 9) by applying the same analysis. Note that the computational complexity of the analytic L^2 distance is quadratic in $\max(M,N)$, that is, irrespective of the number of dimensions. This is markedly lower than the $\mathcal{O}(G^m S)$ complexity for the general KDE-based solution described in Section 5.2.

6. Experiments

The previous sections argued that there are several possible approaches for finding subgroups in data with numeric targets. This section will answer the question how these approaches compare in terms of the subgroups that are found, and will explore the influence of parameter settings on these results.

The most important question (**Q1**) that is raised in this work is whether the use of quality measures that consider complete distributions allows for the discovery of subgroups that are overlooked using simple statistics. This question is answered in Section 6.2 by comparing the (rankings of) subgroups found using the traditional z-score with those found using the H^2 score, on data with a single target attribute.

A claimed benefit of the density estimation methods is that they can also be applied to data with multiple target attributes. Section 6.3 investigates to what extent relevant subgroups, showing interesting deviations in distribution over the targets, can be discovered by means of the two measures H^2 and L^2 (**Q2**). Additionally, it analyses to what extent the two measures capture the same notion of deviation.

For the purpose of measuring differences over complete distributions, two methods to estimate probability distributions are suggested: one based on histograms and one based on KDE. The former is less complex, but also less robust. So, to assess whether it is a practical alternative in terms of result quality, the third question (**Q3**) is whether these different methods yield similar results. Section 6.4 answers this question by comparing the (rankings of) subgroups found, while using the same quality measure.

Finally, both the KDE and the histogram-based methods have an important parameter that needs to be set: the bandwidth parameter and the number of bins, respectively. The use of Silverman's rule was proposed to determine the bandwidth of the kernel. To determine the number of bins, the use of Silverman's rule is proposed as well; indeed, this rule is used in the experiments that answer questions Q1–Q3. The final question (**Q4**), addressed in Section 6.5, is to ascertain what the impact of this heuristic on the results is.

The experiments were performed using the generic Subgroup Discovery tool Cortana[3] [20, 29], that was extended to incorporate the new approach. To avoid the creation of degenerate covariance matrices[4] (see Section 5), a minimum and maximum coverage was enforced on the size of the subgroup (and therefore its complement in cases where this is relevant). They were set to 5% and 95% of the number of records in the dataset, respectively. Further, a beam search strategy was employed, using a beam size of 100, and for numeric attributes, local discretisation [18] determined the 7 bin boundaries used for the descriptions that could (potentially) all be included in the beam (Cortana's default values). The description language uses the '$=$' operator for nominal attributes, and '$\leq$' and '$\geq$' for numeric attributes. Depending on the setting at hand, a subgroup is compared to either the dataset as a whole or just its complement. This will be duly noted.

6.1. Datasets

Characteristics of the various datasets used in the experiments are presented in Table 1. Although not every dataset is used in every experimental setting, they are listed upfront in order not to repeat information. All datasets have at least one numeric target. Datasets for which multiple numeric attributes are used as targets are mentioned where applicable. The following datasets were used. The first three of these were obtained from the UCI repository [43]. Here, N refers to number of records, C is the cardinality of the target and $\mu_{\mathscr{D}}$ is the target mean. Missing values do not occur for any of the targets, and only for some description attributes of *Adult* (3 out of 14) and *Communities* (15 out of 127).

[3] Available from datamining.liacs.nl/cortana.html.

[4] No such situations were subsequently encountered in any of the experiments.

Note that the cardinality of the target is often (much) lower than the number of records, an observation that holds true for many real-world datasets.

Table 1. Dataset used in the various experiments. In some cases, multiple target attributes simultaneously serve as multidimensional target. Here, N refers to number of records, k is the number of description attributes, C is the cardinality of the target and $\mu_{\mathscr{D}}$ is the target mean.

Dataset	target attribute	N	k	C	$\mu_{\mathscr{D}}$
Waterquality	std_temp	1 060	29	217	2.46
Adult	Age	48 842	14	74	38.64
Communities	ViolentCrimesPerPop	1 994	127	98	0.24
Boston Housing	CRIM	506	12	504	3.61
	MEDV			229	22.53
Mammals	latitude	2 221	124	271	50.83
	longitude			1 659	12.04

The *Waterquality* dataset concerns biodiversity in Slovenian rivers. In the original work [44], 16 attributes, concerning physical characteristics of the water, were (independently) used as target. The remaining 14 attributes give the degree of presence of various types of bacteria, and are used as description attributes. In the experiments below, one of the original 16 target attributes, *std_temp*, is again used as target. All other attributes can be used to build descriptions.

The *Adult* dataset is originally used in a binary classification task to determine whether a person makes over $50 000 a year, using parameters like age and education. When used for regression tasks, the numeric attribute *Age* is an obvious target.

The *Communities* dataset combines socio-economic data, law enforcement data and crime data from communities within the United States. As suggested by the dataset provider, attribute *ViolentCrimesPerPop* is used as target. For each community, it indicates the total number of violent crimes per 100k population.

The *Boston Housing* dataset concerns housing values in suburbs of Boston (MA, USA). The target attributes are *CRIM* (per capita crime rate by town) and *MEDV* (median value of owner-occupied homes in $1 000's). As a crime-related attribute is used for the previous dataset, the same is done here by selecting the former, the latter is suggested by the dataset provider.

The *Mammals* dataset [45, 46] contains information about the presence of various types of mammals at geographical locations in Europe. For the experiments, the target attributes were the *latitude* and *longitude* of the location in Europe. The binary description attributes indicate the presence of each mammal considered. The original dataset also contains numeric information about the climate of each location, but, as the interest is only in the (relative) occurrence of species, and not in climatological circumstances, the experiments ignore this information.

In all figures depicting distributions, the subgroup PDF is plotted in black, whereas the grey line shows either the dataset or complement PDF. Correlation plots are used to compare rankings of results, and include R and τ, to denote the size of the rankings and Kendall's tau rank correlation coefficient, respectively.

6.2. Single Target Experiments: z-score versus H^2 (Q1)

This section addressed research question Q1, to see whether the use of quality measures that consider complete distributions allows for the discovery of subgroups that are over-

looked using simple statistics. This is done by checking whether the full distribution measure H^2 ranks high subgroups that are ranked low by z-score. The discussion of the results is broadened a little, as it also looks at what subgroups are ranked high by z-score, but low by H^2. This allows for a better insight into the difference of the two measures.

For these experiments, the *Waterquality* and *Adult* regression datasets were used. Except for the quality measure, all search parameters were the same over the experiments. The main interest is in the comparison of the result sets, or rankings, obtained in the experiments using the $|z\text{-score}|$ and the squared Hellinger distance (H^2). Note that the absolute z-score ($|z\text{-score}|$) is used, as a subgroup could be interesting both with a surprisingly high, as well as low mean value for the target. Also, the H^2 distance has no preference for any section at which the distributions differ most. As the original z-score contrasts subgroups to the whole dataset, the same strategy is followed with the H^2 measure, of which the continuous variant H^2_c is used. Each table shows the subgroup description, the size of the subgroup (extension), and the rank of the subgroup. For reasons of presentation, column names read like 'H^2' instead of 'rank using H^2'. Thus, the numbers listed are *not* quality scores of the subgroups, but their rank.

6.2.1. Waterquality

The first is a simple experiment on the *Waterquality* dataset, with target attribute *std_temp*, considering subgroups of depth 1 (only a single condition). Considering only depth 1 subgroups avoids any beam selection effects, causing the measures to produce results with the exact same subgroups, allowing a direct comparison in terms of the two produced rankings. The size of the result set R is 199, meaning 199 subgroups are found within the search constraints. Table 2 lists the top-10 subgroups ranked by H^2 (left) and $|z\text{-score}|$ (right), and the rank these subgroups attained in the ranking produced by the other measure. The top-ranking subgroups show a distribution where the mean is substantially different from that of the overall dataset. Hence, these subgroups appear high in the rankings of both H^2 and $|z\text{-score}|$. Also, five subgroups feature in both top-10s, indicating that there exists some general agreement between the measures about what subgroups are interesting. However, moving slightly down the list of H^2 subgroups, one encounters at rank 22 a subgroup that shows a bimodal distribution within the subgroup that is much different from the unimodal distribution of the dataset. The corresponding description is '$sio2 \geq 2.3217$', and the subgroup has a size of 134, or 12.6% of the data. This interesting subgroup appears among the very worst of the $|z\text{-score}|$-based ranking, at rank 188 of the 199 results found. The low interestingness for $|z\text{-score}|$ of this subgroup is caused by the relative small difference in mean: $\mu_s = 2.508$ compared to $\mu_{\mathscr{D}} = 2.464$ (standard deviation $\sigma_{\mathscr{D}} = 1.060$). In other words, this subgroup ranks fairly high on the proposed quality measure H^2, whereas it would have been completely overlook and judged uninteresting with this traditional quality measure. This supports a positive answer to research question Q1. The distribution of this subgroup already featured in Figure 1 in the introduction of this work.

Next, consider Figure 4, which shows the rank correlation plot related to these two rankings of results. When two rankings are very similar, the points in such a plot will all lie close to the depicted diagonal, as subgroups attain comparable ranks in both experiments. The general trend is indeed along the diagonal, indicating that both measures generally rate subgroups similarly. However, in the lower right corner of the diagram, subgroups appear that are interesting according to H^2 but not according to $|z\text{-score}|$. The

Table 2. Experiments on dataset *Waterquality*, using target attribute *std_temp* and a search depth of 1. The same 199 subgroups (R = 199) are found in both experiments. Shown here are the top-10 subgroups as ranked by H^2 (left) and $|z$-score$|$ (right). Each line shows the subgroup description, size n, and the rank the subgroup attained in the H^2 and $|z$-score$|$ experiment. Five subgroups feature in both top-10s, indicating that there exists some general agreement between the measures about what subgroups are interesting. Also, there are no subgroups that rank high on one ranking, and low on the other.

Description	n	H^2	$\|z\|$
no2 $\leq$ 0.0	137	1	18
o2 $\geq$ 5.3437057	133	2	12
no2 $\geq$ 0.80609417	140	3	7
o2 $\leq$ 3.313926	136	4	9
o2sat $\geq$ 5.394927	133	5	8
k2cr2o7 $\leq$ 0.13867278	134	6	20
o2 $\leq$ 3.8938632	270	7	1
kmno4 $\leq$ 0.20269547	144	8	25
o2 $\geq$ 4.970889	277	9	2
kmno4 $\leq$ 0.2567476	318	10	15

Description	n	H^2	$\|z\|$
o2 $\leq$ 3.8938632	270	7	1
o2 $\geq$ 4.970889	277	9	2
o2 $\geq$ 4.722345	412	14	3
o2 $\leq$ 4.225256	414	13	4
o2 $\geq$ 4.4738	551	32	5
o2 $\leq$ 4.4738	533	23	6
no2 $\geq$ 0.80609417	140	3	7
o2sat $\geq$ 5.394927	133	5	8
o2 $\leq$ 3.313926	136	4	9
conduct $\geq$ 3.567889	265	21	10

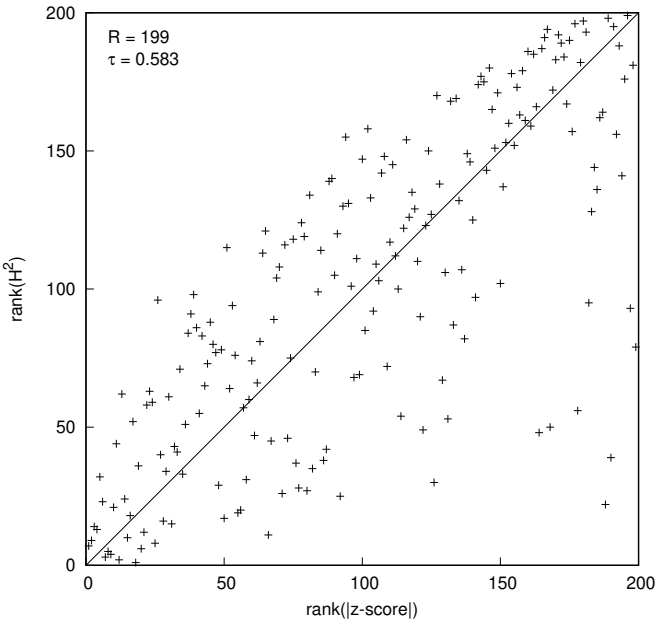


Figure 4. Rank correlation plot for results obtained on the *Waterquality* dataset utilising quality measures $|z$-score$|$ and H^2. The general trend is along the diagonal, indicating that both measures generally rate subgroups similarly. However, some subgroups rank much higher for H^2 than for $|z$-score$|$. Noteworthy is the example with rank 22 for H^2 and 188 for $|z$-score$|$, its PDF is shown in Figure 1. The rankings have a Kendall's tau correlation of 0.583.

most extreme subgroup in the lower right corner is the one discussed in the paragraph above, but several more examples can be recognised in the plot.

Conversely, the complete absence of subgroups in the upper left corner is equally noteworthy. It indicates that there was not a single subgroup that occurred high in the $|z$-score$|$ ranking and at the bottom of the H^2 ranking. So, $|z$-score$|$ does not deem anything

Table 3. A depth 1 search for dataset *Adult*, using target attribute *Age*, yields 63 subgroups. With six subgroups in both top-10s, the rankings generally agree with each other, but for some subgroups the ranking difference is quite large. Noteworthy examples are those that rank 2, 4 and 5 for H^2. For reasons of presentation, attribute name *Marital status* is changed to *Marital s.*.

Description	n	H^2	$\lvert z\rvert$	Description	n	H^2	$\lvert z\rvert$
Relationship = 'Own-child'	7581	1	2	Marital s. = 'Never-married'	16117	3	1
Hours per week $\leq$ 28.0	6136	2	26	Relationship = 'Own-child'	7581	1	2
Marital s. = 'Never-married'	16117	3	1	Relationship = 'Husband'	19716	9	3
Occupation = '?'	2809	4	44	Marital s. = 'Married-civ-spouse'	22379	10	4
Workclass = '?'	2799	5	43	Class = '>50K'	11687	6	5
Class = '>50K'	11687	6	5	Workclass = 'Self-emp-not-inc'	3862	11	6
Education = 'Masters'	2657	7	12	Marital s. = 'Divorced'	6633	8	7
Marital s. = 'Divorced'	6633	8	7	Class = '≤50K'	37155	28	8
Relationship = 'Husband'	19716	9	3	Workclass = 'Private'	33906	37	9
Marital s. = 'Married-civ-spouse'	22379	10	4	Education = 'Some-college'	10878	24	10

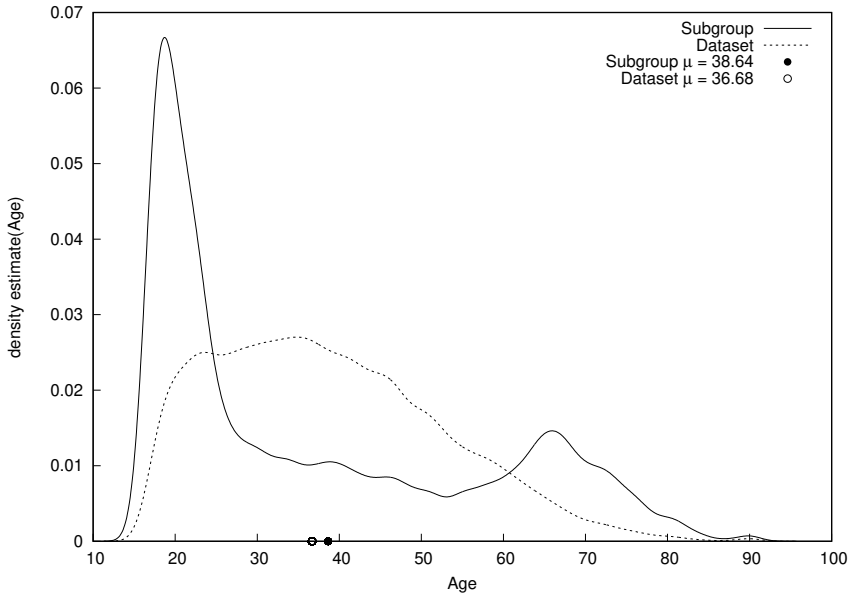


Figure 5. The second-best result obtained on the *Adult* dataset using H^2: '*Hours per week* $\leq$ *28*'. This subgroup covers 6 136 records, or 12.6% of the dataset, and is ranked on place 26 of 63 results for quality measure $|z$-score$|$. This is the result of the modest difference in mean.

interesting that is missed by H^2.

6.2.2. Adult

The *Adult* dataset describes over 48 thousand Americans, and in this context, the interest is in the attribute *Age*. This target attribute shows a unimodal distribution with a mean of $\mu_{\mathscr{D}} = 38.64$ years. Again, the proposed quality measure is compared with the traditional measure for numeric targets, and the focus is on subgroups that rank high on H^2 and relatively low on $|z$-score$|$.

Table 3 shows the results. The subgroup '*Hours per week* $\leq$ *28*' ranks second for H^2 and is a good example of a subgroup found by the proposed measure that is deemed

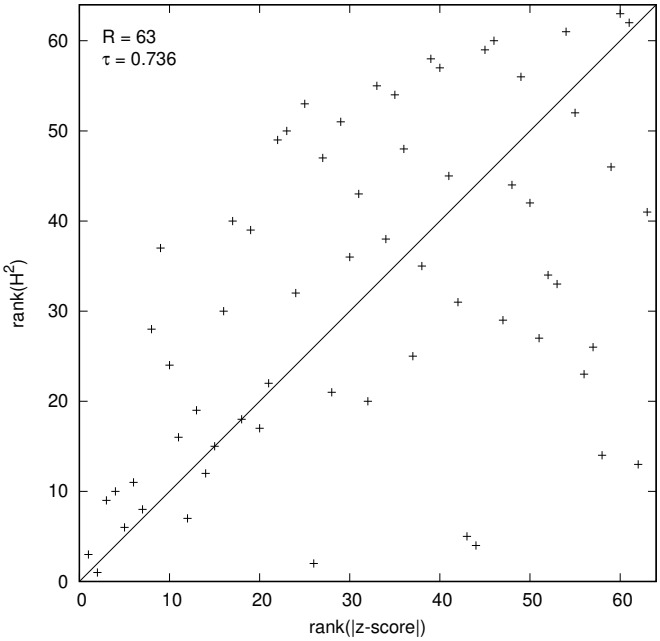


Figure 6. Rank correlation plot for results obtained on the *Adult* dataset utilising quality measures $|z\text{-score}|$ and H^2. Noteworthy examples are those that rank 2, 4, and 5 for H^2, they rank 26, 44, and 43 for $|z\text{-score}|$, respectively. The rankings have a Kendall's tau correlation of 0.736.

uninteresting (rank 26 out of 63) using $|z\text{-score}|$. Figure 5 shows a comparison of this subgroup to the overall distribution. Though less extreme, it is in fact similar to that in Figure 1, showing distributions that are bimodal in the subgroup, and unimodal in the data. The subgroup selects both surprisingly low and high age groups, whereas the average age within this subgroup is not substantially different from that of the dataset (36.68 years compared to $\mu_{\mathscr{D}} = 38.64$, standard deviation $\sigma_{\mathscr{D}} = 13.71$). Because of this small difference, $|z\text{-score}|$ would not consider the subgroup interesting. The H^2 measure does label the subgroup as interesting, as the distributions as a whole differ substantially. Other noteworthy examples include the subgroups that rank 4 and 5 for H^2, they list 44 and 43 for $|z\text{-score}|$, respectively.

Finally, when considering the rank correlation plot for the *Adult* dataset (Figure 6), the same trend is visible as for the *Waterquality* dataset. Generally, points fall along the diagonal, resulting in an even higher tau score of 0.736 for this dataset, compared to the 0.538 for *Waterquality*. And, as before, the left upper corner is empty, whereas the right lower section is not. This again indicates that some subgroups score high on H^2, but not on $|z\text{-score}|$, and that there are no subgroups that are ranked high by $|z\text{-score}|$ that are considered uninteresting by H^2.

Figure 7 shows the rank correlation plots for the *Waterquality* (left) and *Adult* (right) datasets, as discussed above, but with a search depth of 2, allowing for conjunctions of at most two conditions. The same trend is visible as for the depth 1 results, the upper left corner is empty, the lower right corner is not.

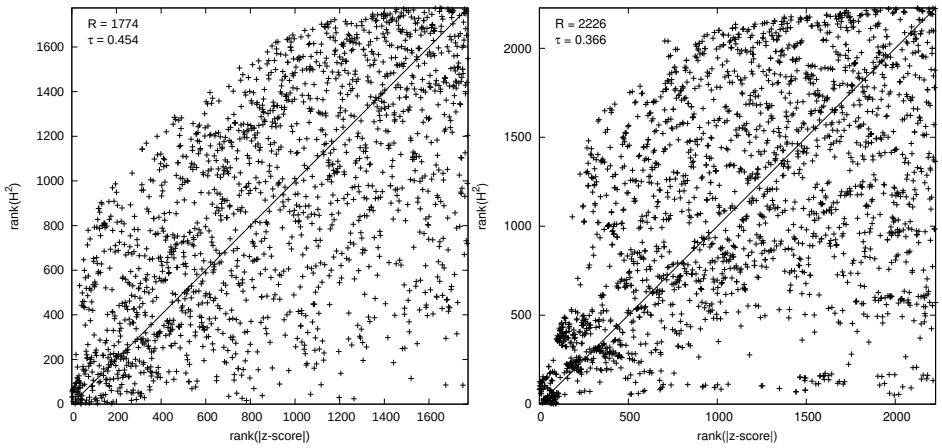


Figure 7. Rank correlation plots for results obtained on the *Waterquality* (left) and *Adult* (right) datasets, utilising quality measures $|z\text{-score}|$ and H^2, and a search depth of 2. The *Waterquality* and *Adult* rankings have a Kendall's tau correlation of 0.454 and 0.366, respectively. As observed in the depth 1 results, the upper left corner is empty, the lower right corner is not. (To produce comparable rankings, an exhaustive (depth-first) search was used, instead of a beam search.)

6.2.3. Communities

Finally, a third set of experiments was performed on the one remaining single-target dataset, *Communities*. The ranks of the two measures are rather correlated ($\tau = 0.736$), suggesting that this dataset contains only few subgroups found by H^2 while being overlooked by more traditional measures (see Figure 8). Still, in the lower left-hand corner of the plot, one can observe several subgroups that appear at the top of the ranking for one measure, but not quite so for the other. Specifically below the diagonal, near the x-axis, there is a group of high-ranking subgroups for H^2 that only appear around rank 200 for $|z\text{-score}|$. Figure 9 presents one such subgroup ('$PctKids2Par \leq 0.86$'), which has ranks 5 for H^2 and rank 218 for $|z\text{-score}|$. Even though a rank-218 subgroup is within the best 20% of subgroups, one could argue that below rank 10, subgroups are unlikely to be inspected by a human analyst. This subgroups is primarily interesting because it has a substantially lower average (so not the multi-modal distribution that H^2 particularly excels in), which is also recognised by $|z\text{-score}|$. However the latter includes a factor for the size of the subgroup, which is this case is only 12.8%. This experiment demonstrates that H^2 doesn't necessarily find exceptional and previously overlooked subgroups in all datasets, and it really depends on the data at hand whether deviating multi-modal distributions can be discovered.

6.3. Discovering Meaningful Subgroups over Multiple Targets (Q2)

This section shows some results of experiments that used a multidimensional target. It addresses research question Q2 by investigating how the two density distance measures H^2 and L^2 perform in this setting. As in the previous section, the comparison is done by way of rank correlation plots. However, subgroups are compared to their complement, instead of the overall dataset, and therefore only operator '$\geq$' is used. The aim is to get a better understanding of the different behaviours of the two measures, and gauge to

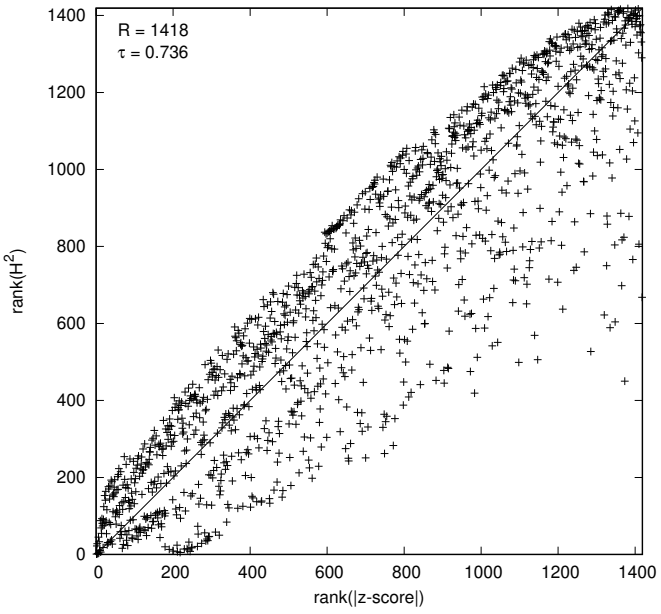


Figure 8. Rank correlation plot for *Communities*, showing considerable correlation between the two measures $|z\text{-score}|$ and H^2. Interesting subgroups discovered by H^2 that are valued only moderately by $|z\text{-score}|$ appear in the lower left corner, below the diagonal near the x-axis (see Figure 9).

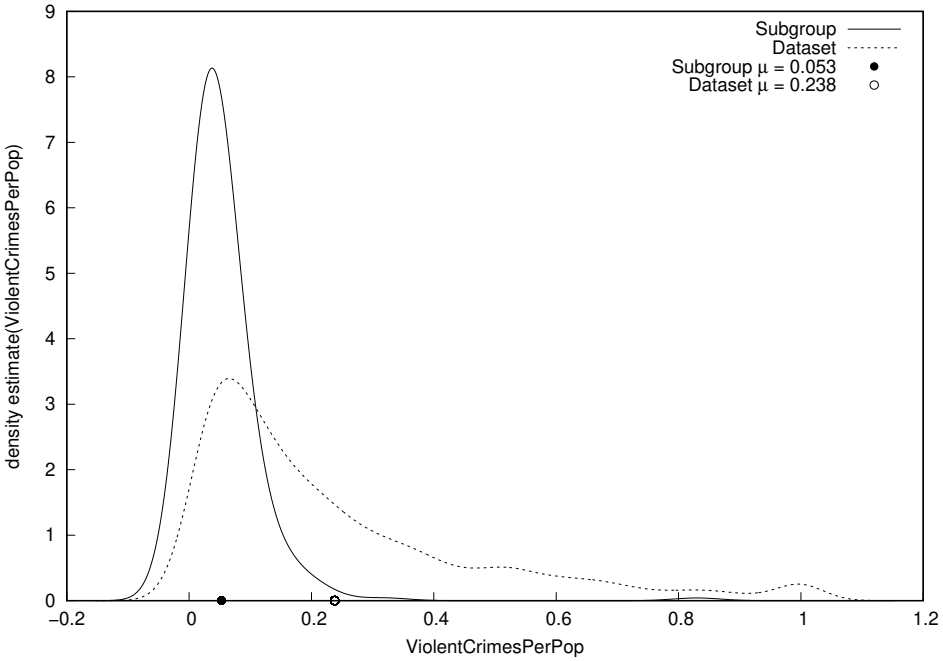


Figure 9. Example subgroup that ranks high in H^2 (5), but not quite so for $|z\text{-score}|$ (218), for the *Communities* dataset. Even though there is a substantial difference in mean between the subgroup and the data, the $|z\text{-score}|$ measure does not rank this subgroup higher due to its relatively small size (12.8% of the data).

Table 4. For dataset *Mammals*, using attributes *latitude* and *longitude* as two-dimensional target, 67 subgroups are found at depth 1. Shown here are the top-10 subgroups as ranked by H^2 (left) and L^2 (right). Invariably, subgroups ranked high by H^2 rank low with L^2. Also, on average, subgroups in the H^2 top-10 ranking are much smaller. For reasons of presentation, all subgroup descriptions are changed from '*species* = *1*' to '*species*'.

Description	n	H^2	L^2	Description	n	H^2	L^2
Talpa occidentalis	128	1	66	Dama dama	663	63	1
Lepus granatensis	150	2	61	Cervus elaphus	1075	62	2
Clethrionomys rufocanus	219	3	67	Lutra lutra	1364	65	3
Genetta genetta	285	4	59	Neomys anomalus	530	46	4
Microtus duodecimcostatus	203	5	56	Ovis ammon	336	51	5
Lemmus lemmus	169	6	65	Crocidura leucodon	584	33	6
Alces alces	653	7	64	Felis silvestris	630	47	7
Sorex caecutiens	170	8	62	Arvicola terrestris	1346	60	8
Gulo gulo	161	9	63	Crocidura suaveolens	663	37	9
Arvicola sapidus	372	10	57	Glis glis	712	40	10

what extent they capture the same information. Additionally, it demonstrates the ability to discover relevant findings by some illustrative example subgroups.

Also, as the L^2 measure allows for an analytical solution, ensuring a low computational complexity, the difference in computation time is quantified.

6.3.1. Mammals

For the first experiments, the *Mammals* dataset was used. In this two-dimensional setting, the target was formed by the attributes *latitude* and *longitude*. The attributes indicating the presence or absence of a different species were used as description attributes.

Figure 10 shows a rank comparison of the results obtained using the H^2 and L^2 measures. It is instantly clear that there is very little positive correlation between the two rankings. In fact, some of the best scoring subgroups for one measure rank among the worst for the other, leading to a tau score of -0.566.

Besides the rank correlation plot, the top-10 results are shown for each measure in Table 4. Here, there are 67 subgroups within the search constraints. Again, note that subgroups that rank very high using one measure rank low for the other. For example, the subgroups ranked first and third using the H^2 measure (see Table 4 left), ranked second last and last using the L^2 measure. When considering the top-10 reported for L^2, the same behaviour is observed, though it appears to be a little less extreme, as there are some subgroups that occur halfway through the H^2 ranking. Noteworthy also is the difference in size of the high-ranking subgroups. Where the average size for the top-10 H^2 subgroups is 251, it is 790 for the L^2 top-10. This is 11.3% and 35.6% of the data, respectively. So, for this dataset, it seems H^2 favours smaller subgroups more, and L^2 has a preference for larger ones. Section 7 will come back to this observation.

6.3.2. Demonstration of Results

Figure 11 shows the top result of the H^2 ranking: *Talpa occidentalis*, also know as the Spanish mole. This subgroup covers 128 records, or 5.8% of the dataset. The plot shows the difference between the PDFs – a distribution over latitude and longitude – of the subgroup and the complement as a greyscale. The bright region corresponds to the area where the values for subgroup PDF are high, and those of the complement are low, and

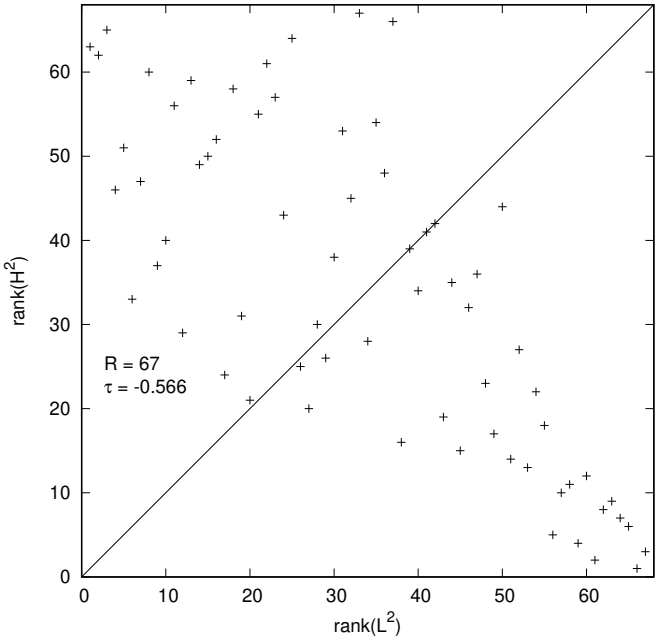


Figure 10. Rank correlation plot for *Mammals*, using a two-dimensional target, showing the unexpected inverse relation between the two measures H^2 and L^2 ($\tau = -0.566$) on this dataset.

thus where Spanish moles can be expected. This subgroup ranks highest, because it is the species that is most extreme in its distribution throughout Europe. Note that such a subgroup could also have been found by a z-score-like measure that computes the distance of centre of mass of the subgroup to that of the centre of Europe. As such, it demonstrates that species with unusual distributions can be found by H^2, but it is not an example of a multimodal distribution.

Figure 12 shows a more interesting result in this respect. This result was ranked as number 11 for H^2, and relates to *Crocidura Russula*, better known as the greater white-toothed shrew ($n = 582$, 26.2% of the dataset). Again, the result indicates that a certain species occurs much more frequently in some distinct parts of Europe than elsewhere. More importantly, the distribution of this shrew is not in a corner of Europe, but rather is located in a fairly central region. Both on the North-West, North, and East, the shrew is less-likely to be found, and then there are those areas (such as Scandinavia and Italy, where the presence is moderate). Especially when considering the complement (the surrounding darker horse shoe-shaped region), such a subgroup would be impossible to find with simpler measures in the vain of z-score.

Finally, Figure 13 demonstrates the distribution of *Dama dama* (fallow deer), the top-ranking result for the L^2 measure ($n = 663$, 29.9% of the dataset). This results is also relevant, since it shows two modes where the dear is particularly common (the British Isles and Eastern Europe), and two regions where it is not (Southern France and North-Western Russia and Finland). Additionally, this result is a testament to the power of a multivariate (Exceptional Model Mining) approach, as it is an XOR problem, that is, this subgroup would not be identified using either *latitude* or *longitude* alone.



Figure 11. Top result of the H^2 ranking, listed in Table 4, for the *Mammals* dataset. The heatmap shows the difference between the PDFs of the subgroup and the complement of *Talpa occidentalis* (Spanish mole). The bright region indicates that in that part of Europe this species occurs more frequently than in the rest of Europe. A simple centrality-based measure might also find this result, as this region lies far from the centre of Europe.

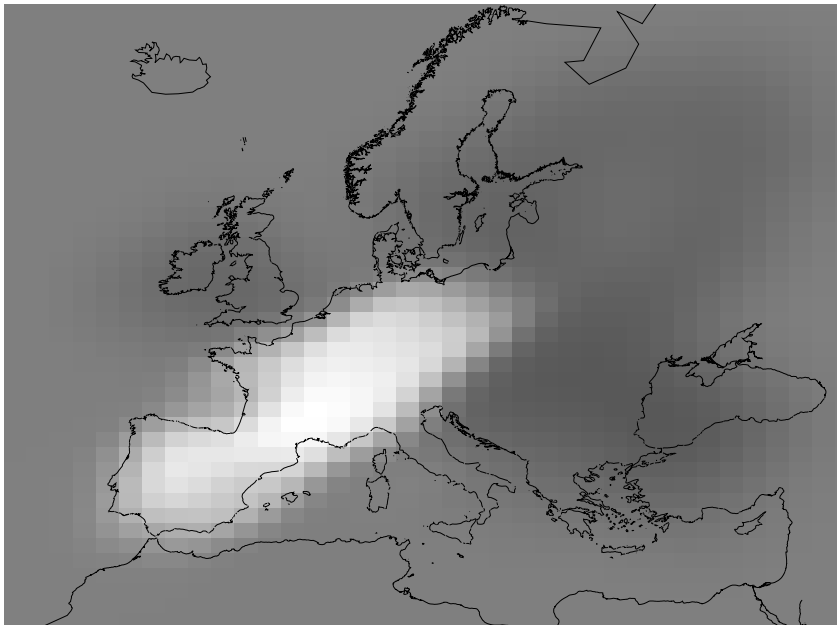


Figure 12. This result ranked 11^{th} in the H^2 ranking, and involves *Crocidura russula* (greater white-toothed shrew). This species of mouse occurs more often in a fairly central region, and a subgroup like this would be impossible to find with simpler measures in the vain of z-score.

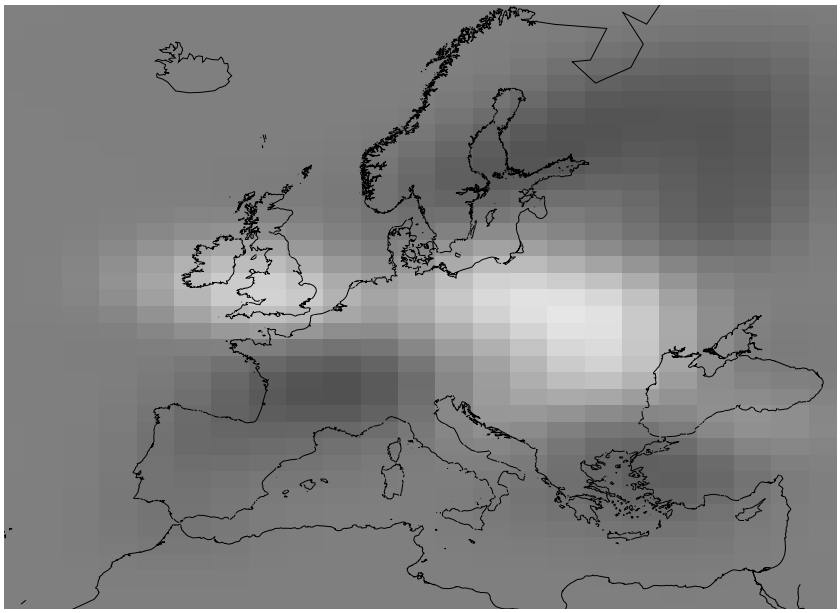


Figure 13. For *Mammals*, this is the top-ranking result in the L^2 ranking. This species, *Dama Dama* (fallow deer), occurs more frequently in the bright regions than in the rest of Europe. This is an exciting higher-dimensional, multimodal result, that shows that the proposed method goes beyond 'classical' centre of mass deviations when searching for interesting subgroups.

In conclusion, both H^2 and L^2 are capable of identifying interesting subgroups in this dataset with a two-dimensional target. The results show both 'classical' results where merely the centre of mass of the subgroup deviates, as well as attractive multimodal results. Although the rank correlation (with a tau score of -0.566) indicates that the two measures favour subgroups in this dataset differently, they still both produce relevant findings.

6.3.3. Boston Housing

The *Boston Housing* dataset concerns housing values in suburbs of Boston (MA, USA), and the attributes *CRIM* (per capita crime rate by town) and *MEDV* (median value of owner-occupied homes in \$1000's) were designated as target. Figure 14 shows the rank correlation of the results obtained using the H^2 and L^2 measures. Clearly, there is not much correlation between the two measures (with a tau score of 0.267). There are four subgroups (involving in total two attributes) that appear in both top-10s, but there are also several examples of top subgroups in one ranking that appear in the lower half of the competing ranking. This, again, shows that the quality measures gauge the quality of subgroups in alternative manners.

Figure 15 shows the top-ranking result for the H^2 measure. This subgroup occurs only halfway through the L^2 ranking, at rank 33.

Unlike the *Mammals* results, here L^2 does not show a preference for larger subgroups than H^2. In fact, on average within the top-10, the subgroups are somewhat smaller.

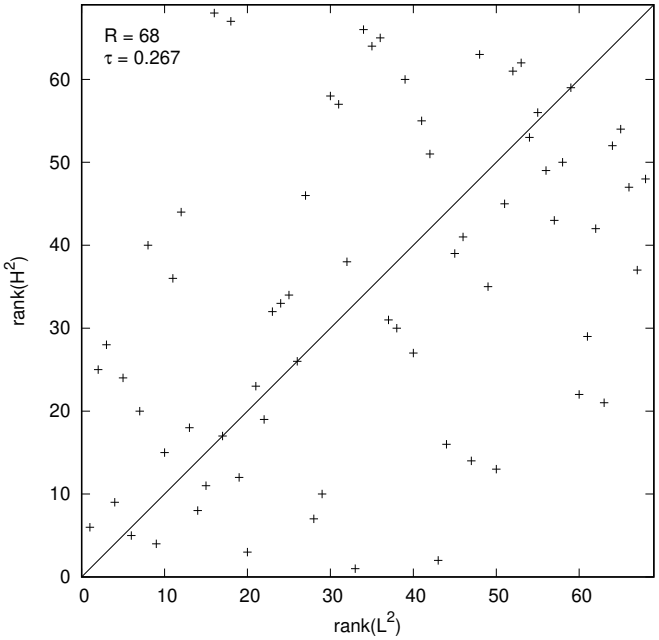


Figure 14. Rank correlation plot for the *Boston Housing* with double numeric targets, demonstrating the measures H^2 and L^2 rate the various subgroups decidedly different ($\tau = 0.267$).

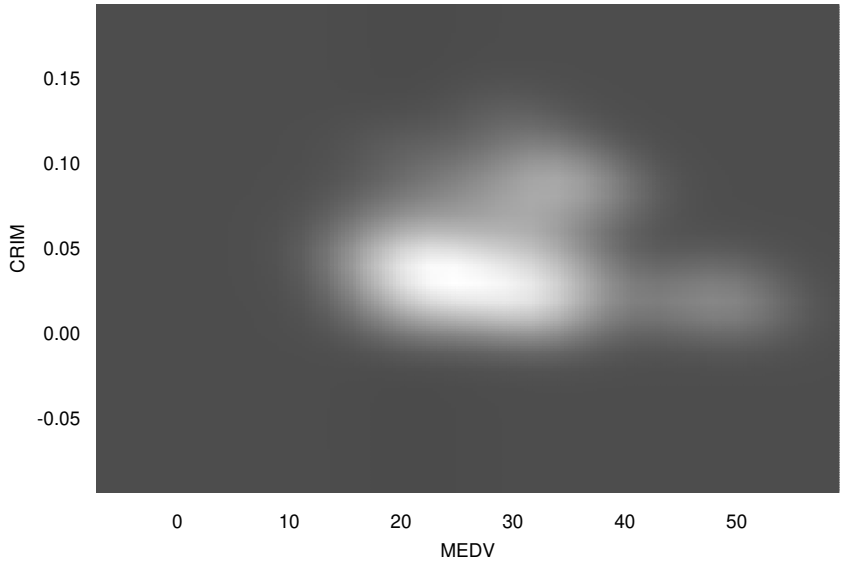


Figure 15. For *Boston Housing*, the top-ranking result in the H^2 ranking attains rank 33 (out of 68) using L^2.

6.3.4. Run times

The final aspect that is considered for Q2 relates to the computational complexity of the H^2 and L^2 measures, as described in Sections 5.2 and 5.3, respectively. Table 5 shows run

Table 5. Run times for the *Mammals* and *Boston Housing* datasets for two-dimensional target experiments, comparing a naive grid-based H^2 implementation, and the analytical L^2 solution at various depths. Run times are given in minutes and seconds (m:ss). For H^2, the granularity of the grid, controlled by the variance of the subgroup, impacts run times more than the number of examined subgroup candidates. For these experiments, run times for L^2 are 5 to 300 times less than that of H^2.

Dataset	targets	depth	number of candidates	H^2	L^2	$\frac{L^2}{H^2}$
Mammals	latitude/longitude	1	245	0:15	0:03	0.200
		2	14 823	10:14	1:48	0.179
		3	578 521	348:22	50:02	0.144
Boston Housing	CRIM/MEDV	1	69 1:55	<0:01	–	
		2	4 485	65:53	0:10	0.003
		3	252 947	2497:19	6:42	0.003

Table 6. Run times for the *Boston Housing* dataset, using the L^2 measure, for targets consisting of two, four and eight attributes. Run times do not dramatically increase with the increasing target dimension, indicating that L^2 scales well, and can be used for high-dimensional targets.

Targets	depth	number of candidates	run time
CRIM/ZN	1	34	<0:01
	2	1 157	0:02
	3	33 697	0:58
CRIM/ZN/INDUX/NOX	1	34	<0:01
	2	1 157	0:03
	3	33 697	1:10
CRIM/ZN/INDUX/NOX/RM/AGE/DIS/RAD	1	34	<0:01
	2	1 157	0:05
	3	33 697	2:00

times for the *Mammals* and *Boston Housing* dataset, for the experiments described above. It becomes clear that for the H^2 implementation, the granularity of the grid impacts run times more than the number of examined subgroup candidates. The *Mammals* dataset has about four times more records than the *Boston Housing* dataset, and more candidate subgroups are tested during equivalent experiments. Yet, run times for the *Mammals* experiments are far lower than those for *Boston Housing*. This is the result of the greater number of grid points used for the latter dataset, caused by a smaller (relative) step size δ. For these experiments, run times for L^2 are 5 to 300 times lower than those of H^2.

Demonstratively, for the *Boston Housing* dataset, additional experiments were performed, using two, four and eight target attributes. These results serve just as proof of the capability of the analytic L^2 approach. There is no clear interpretation of the target used, and visualisation of high-dimensional targets is not straightforward, so no results are discussed other than the run times. To make a fair comparison, the exact same subgroup candidate set should be generated in each setting. In the two-dimensional target setting, the first two numeric attributes in the dataset were designated as targets, the next six numeric attributes were disabled, and the remaining five attributes served as description attributes. In the four-dimensional setting, the first four numeric attributes served as target and the next four were disabled. As can be seen in Table 6, the run times do not dramatically increase with the increasing target dimension.

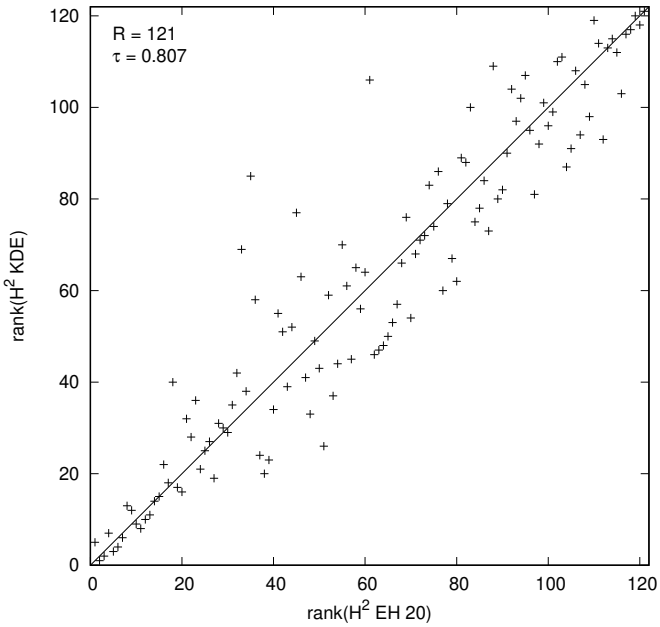


Figure 16. Rank correlation plot for results obtained on the *Waterquality* dataset utilising quality measures performing probability estimation using equal-height (EH) histograms and KDE. Most points lie close to the diagonal, indicating that the two techniques assigned similar ranks to most subgroups. The rankings have a Kendall's tau correlation of 0.807. The most extreme discrepancy is found for subgroup '*Gammarus fossarum* $\geq$ *5.0*', which ranks 35 and 85 (out of 121) for the EH and KDE-based technique, respectively. The probability estimates for this subgroup are shown in Figure 17 (KDE), and Figure 18 (EH).

6.4. Comparison of Density Estimation Techniques (Q3)

This section addresses Q3 and investigate the effects of different methods of performing probability estimation on the (ranking of) the discovered subgroups. The experiments ran to this end were mostly identical in set-up to those described in Section 6.2. A comparison is made between the density estimates produced by equal-height (EH) and equal-width (EW) histograms, and KDE. KDE methods circumvent some of the drawbacks encountered in histogram techniques, but are computationally more complex. Here, this is exacerbated by the fact that, for KDE, bandwidth h is computed per subgroup, whereas for the histograms, the bin boundaries are computed only once, in advance of the experiment. So, if the results of the simple histogram techniques are comparable and cheaper to those of a more complex KDE approach, one might opt for a simpler histogram implementation. The experiments were performed on the *Adult*, *Communities*, and *Waterquality* datasets. Further, subgroups are compared to their complement, instead of to the dataset as a whole.

The first comparison is between EH and KDE, on the *Waterquality* dataset. Figure 16 depicts the rank correlation plot for this experiment. This plot, and those for the other two datasets presented further down, indicates that on the whole, the rankings produced by the two density estimation techniques are fairly consistent. This is evident from the many points close to the diagonal, and also from the high rank correlation coefficient of $\tau = 0.807$. The lack of points in the upper left and lower right corner indicates that there are

no subgroups that are deemed interesting through one estimation, but are not recognised as such by the other approach. However, there are a few points that lie a bit further from the diagonal, representing subgroups that are ranked higher by EH than by KDE. As Figure 16 shows these cases are rare, but they do occur. Consider the most extreme example in this setting, subgroup '*Gammarus fossarum* $\geq$ *5.0*'[5], with rank 35 and 85 (out of 121) for the EH and KDE-based technique, respectively. This subgroup describes the presence (5.0 represents abundance) of a species of *Gammarus*, an amphipod crustacean genus in the family Gammaridae.

Figures 17 and 18 show the probability estimates for this subgroup. Figure 17 shows the PDF obtained using the KDE-based technique. Here, the complement values are more evenly distributed over the *std_temp* domain compared to the subgroup, which has relatively many values concentrated around 2.09. Except for this narrow peak, the two distributions are fairly close to each other.

Exactly why the density estimate produced by the EH approach appears more interesting than suggested by KDE becomes apparent from Figure 18. This histogram indeed confirms that there is a reasonably unusual peak around 2.09, but two bins stand out in particular: $(1.344085, 1.519400]$ and $(2.064826, 2.220661]$. The empty subgroup bin at $(1.344085, 1.519400]$, the fourth bin from the left, suggests that there is an unusual gap in the distribution of the subgroup (the solid black bars) that is not evident in the complement. Indeed, when inspecting the subgroup records, the *std_temp* target values show a gap between 1.2856461 and 1.5388794. Furthermore, when considering bin $(2.064826, 2.220661]$, the one with the tallest bar of all, the estimated subgroup probability is almost 4.5 times higher than that of the complement. And although in the KDE case the density around the peak for the subgroup also markedly differs from the density of the complement, the densities differ only by a factor of two.

Both cases could be caused by over-smoothing in the KDE estimate, something that is known to happen when employing Silverman's rule of thumb on non-normal or skewed distributions [47]. On the other hand, histograms are known to be sensitive to both the number of bins and the bin boundary placement [11]. Figure 3 above already presented an example of this, and a real-world example on the current dataset is discussed in more detail below. Finally, Figure 19 shows rank correlation plots for the two other datasets considered, *Adult* (left) and *Communities* (right). These plots do not show any substantial deviation from the diagonal, and also the rank correlations are high ($\tau = 0.971$ and $\tau = 0.942$).

Next, the robustness of histogram-based techniques is addressed by considering an extreme example of a subgroup ranked much higher by EH than by KDE. This behaviour is caused by an artifact due to unfortunate placement of bin boundaries. For the *Waterquality* dataset, the same KDE technique was used as before, and results were obtained using the EH technique, now using 21 instead of 20 bins. Figure 20 shows the rank correlation for this setting, and one point in the upper left corner stands out in particular. It represents subgroup '*std_ph* $\geq$ *22.8212*', with rank 17 and 106 (out of 121) for the EH and KDE-based technique, respectively. Figures 21 and 22 show the probability estimates for this subgroup.

Figure 21 shows the PDF obtained using the KDE-based technique. Here, the subgroup values are more evenly distributed over the *std_temp* domain compared to the com-

[5]The dataset uses code '49700' for this species. For presentation, the name is used instead.

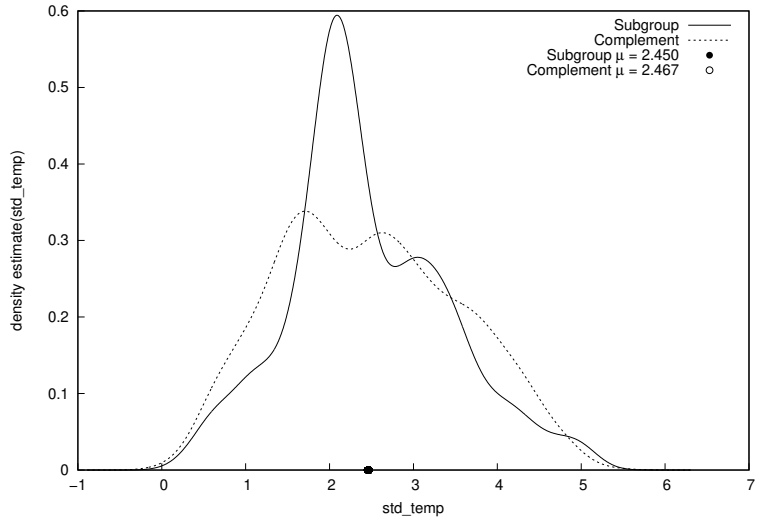


Figure 17. For the subgroup and complement of '*Gammarus fossarum* ≥ 5.0', KDE created PDFs that do not differ much overall, save the narrow peak in the centre.

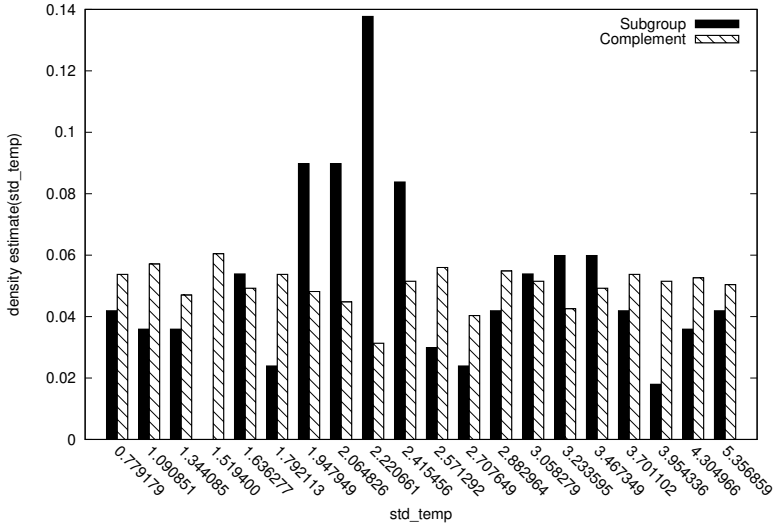


Figure 18. Probability estimate for the subgroup and complement of '*Gammarus fossarum* ≥ 5.0', using EH histograms. In this plot, the *x*-axis labels represent the (inclusive) upper bounds for each bin. The (exclusive) lower bounds is thus formed by the upper bound of the previous bin. The bar heights, probability estimates, in the PMF sometimes differ substantially between the corresponding bins for the two distributions. Besides the four tallest bars for the subgroup, there is another notable example. For bin $(1.344081, 1.519400]$, the probability for the subgroup is 0.0 and much lower than that for the complement, and probabilities for the adjacent subgroup bins, suggesting a sensitivity with respect to bin boundary placement.

plement, which appears to have relatively many values concentrated slightly above 2.1. Except for this narrow peak, the two distributions are fairly close to each other. Exactly why the density estimate produced by the EH approach appears more interesting than

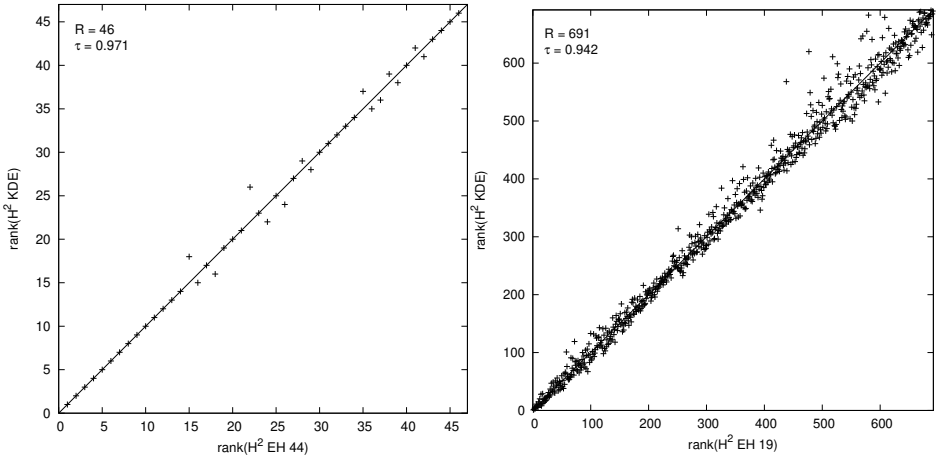


Figure 19. Rank correlation plots for the *Adult* (left) and *Communities* (right) datasets. The ranks are produced employing KDE and EH histograms, using 44 bins in the *Adult* setting, and 19 in the *Communities* setting. In both cases, the rankings do not differ much, yielding Kendall's tau correlations of 0.971 and 0.942, for the *Adult* and *Communities* dataset, respectively.

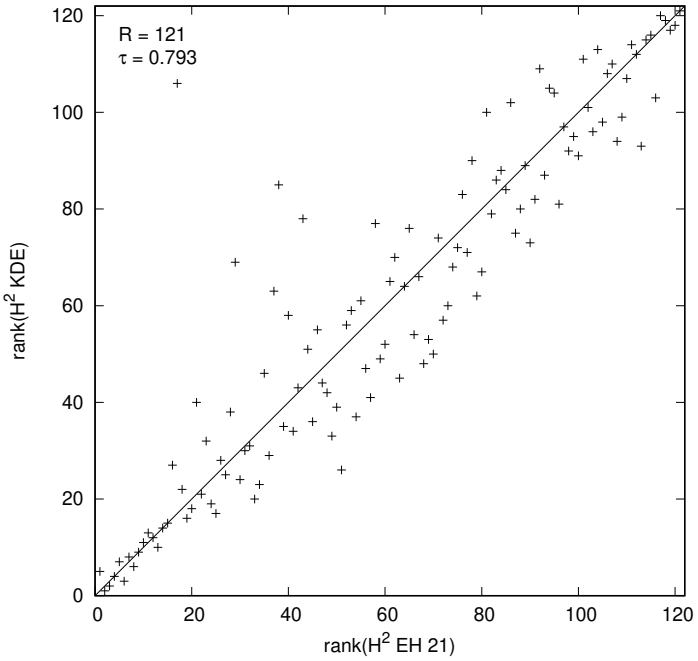


Figure 20. Rank correlation plot for results obtained on the *Waterquality* dataset utilising quality measures performing probability estimation using EH histograms and KDE. The techniques rank most subgroups similarly ($\tau = 0.793$), but some are ranked much higher by EH than KDE, with an extreme example in the upper left corner, the opposite does not occur. The point in the corner represents subgroup '*std_ph* $\geq$ *22.8212*', and its probability estimates are shown in Figures 21 (KDE) and 22 (EH). Further, this plot should be compared to Figure 16, in which EH used 20 bins, compared to the 21 here. The tau scores are similar, 0.807 versus 0.793, but the single added bin results in an extreme outlier, demonstrating issues with robustness.

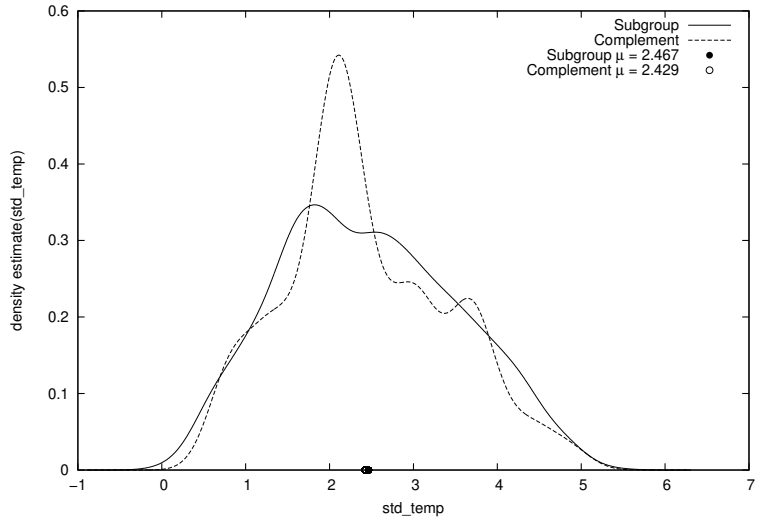


Figure 21. Density estimate using KDE for the subgroup and complement of '*std_pH* $\geq$ *22.8212*'. The PDFs do not differ much overall, save the narrow peak in the centre.

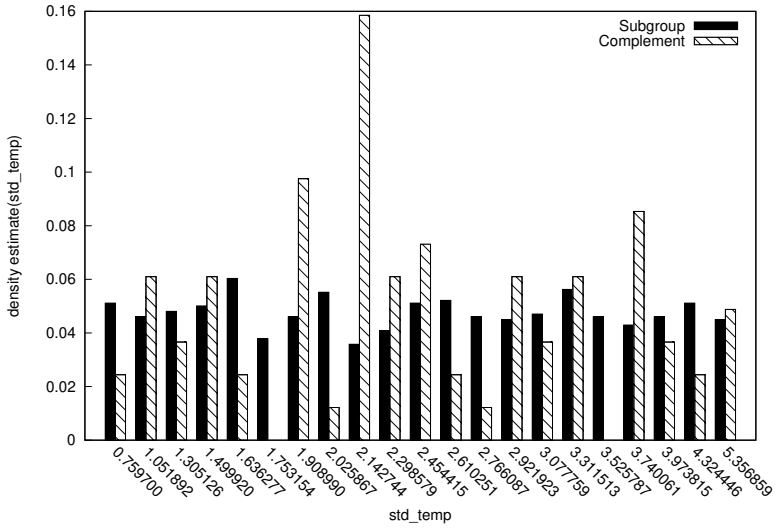


Figure 22. Probability estimate for the subgroup and complement of '*std_temp* $\geq$ *22.8212*', using EH histograms. Like the PDF in Figure 21, this estimate shows a peak in the centre. However, at label 1.753154 and 3.525787, unusual gaps in the distribution of the complement occur, that are not evident for the subgroup. These appear to be artefacts of the boundary (mis)placement.

suggested by KDE becomes apparent from Figure 22. This histogram indeed confirms that there is a reasonably unusual peak around 2.1, but it also suggests that there are two unusual gaps in the distribution of the complement (the striped bars) that are not evident in the subgroup: two empty bins at $(1.636277, 1.753154]$ and $(3.311513, 3.525787]$. Interestingly, in both cases, a neighbouring bin shows a density that is roughly double of what one would expect from the close bins in the subgroup and complement, suggesting

that these empty and 'double' bins are actually an artefact of the boundary placement. More evidence that one is looking at an artefact here rather than a substantive deviation in density comes from the size of the subgroup in question: $n = 978$ versus $n^c = 82$ for the complement, which means 8% of the dataset. Artefacts in density estimates over small sets of data are more likely to occur.

Number of bins When comparing between the two histogram-based estimation methods and KDE, a complicating factor is the choice of the number of histogram bins B. Clearly, if this number is set too low, precision is sacrificed, and very different values will be placed in the same bin. On the other hand, having too many bins will place adjacent values in separate bins, effectively dropping the continuous nature of the attribute. Two very close values that end up in separate bins will be treated essentially the same as two values from either end of the bin range. Figure 23 presents results of experiments where the number of bins is varied, and compares the rank correlation between EH and KDE, and between EW and KDE, respectively. This treats KDE as the golden standard (which may be giving it too much credit), and both EH and EW as the easier-to-compute alternative. The achieved rank correlation (the vertical scale in the three plots) is now an indication of how well an approximation the histogram-based approaches offer in terms of the resulting subgroup rankings.

The three plots do indeed exhibit the expected behaviour, where both EH and EW perform worse when the number of bins drops below a certain number (between 5 and 10 in these datasets). Also, when setting the number of bins too high, a suboptimal rank correlation is achieved. Still, on the whole, EH and EW achieve fairly similar scores, and these scores rank-correlate fairly high with the KDE, as was already observed in the earlier experiments. In the *Waterquality* dataset (top plot), there appears to be a small edge for EW over EH when the number of bins increases. More clearly, there is a drop in performance for EW in *Communities* (bottom plot), when too low a number of bins is selected. This demonstrates a weakness of EW [48] when data has a skew distribution, which is the case for *Communities* but not for the other two datasets. The *ViolentCrimesPerPop* attribute has a long-tailed distribution, with a median of 0.15 and mean of 0.24 (on a domain of [0,1]), and some 65% of the data lie below the mean. The equidistant bin boundary placement of EW leads to many bins in the low density tail, and few in the narrow high density area.

6.5. Silverman's Rule of Thumb (Q4)

Finally, Q4 is addressed by evaluating the effect of the pragmatic choice of using Silverman's rule of thumb to determine the number of bins for the histograms. The earlier Figure 23 shows that using the number of bins suggested by Silverman's rule of thumb does not lead to the absolute highest tau score in any of the experiments. The suggested optimal number of bins is indicated by a dashed vertical line. Table 7 shows the tau scores for the suggested number of bins ($\tau_{silverman}$), and the number of bins leading to the absolute highest tau score ($\tau_{max\ \tau}$). The final column shows how close, relatively, $\tau_{silverman}$ is to $\tau_{max\ \tau}$, by dividing the former by the latter. For both the *Adult* and *Communities* dataset, the scores obtained using the suggested number of bins are close to that of those leading to the absolute highest tau score, especially for the EH technique. For the *Waterquality* dataset, overall tau scores for any number of bins are comparatively low with respect to those obtained for the other datasets. The relative score for the suggested number of bins

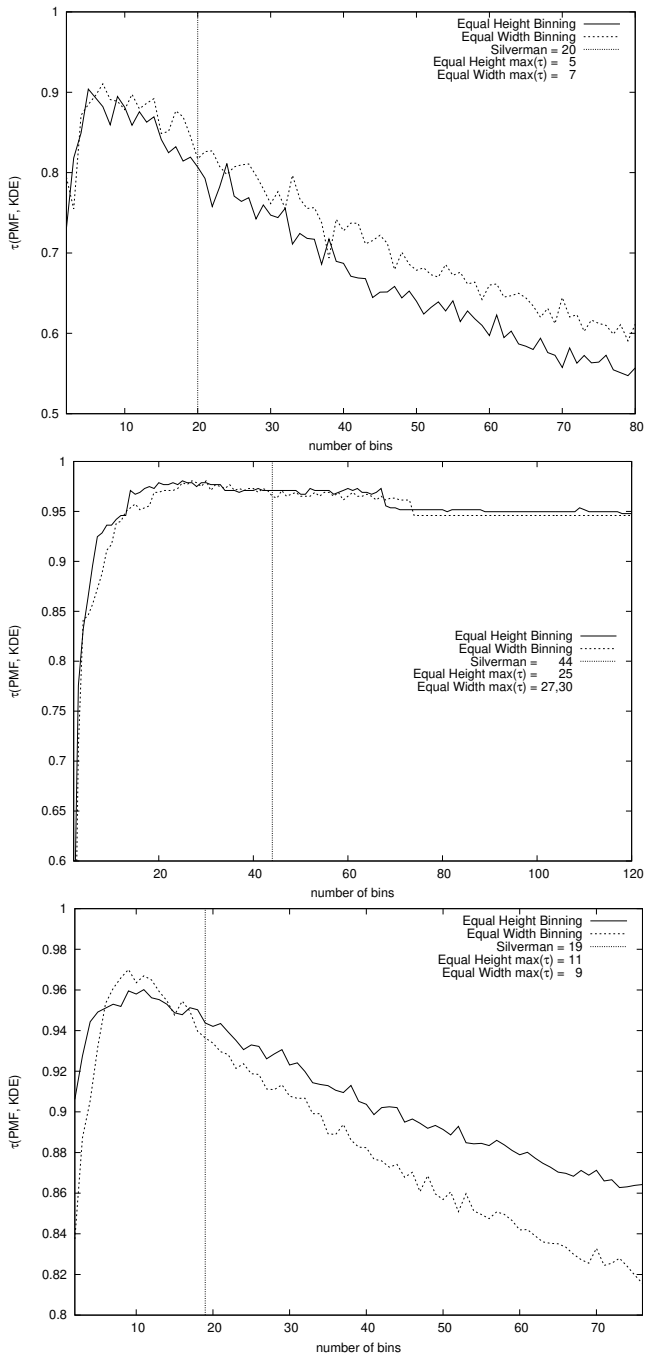


Figure 23. Tau scores for the rankings obtained using EH and EW histograms for the probability estimation versus those obtained using KDE, for the datasets *Waterquality*, *Adult*, and *Communities* (top to bottom). Silverman lists the number of bins suggested by Silverman's rule of thumb, $\max(\tau)$ gives the number of bins resulting in the highest tau score for both histogram types. Silverman's suggestion, the vertical dashed line, is roughly twice as high as $\max(\tau)$.

Table 7. For three datasets, subgroup rankings obtained using equal-height (EH) and equal-width (EW) histograms were compared with the equivalent KDE ranking, resulting in the tau scores listed. For both histogram types, results are given for the setting with the number of bins suggested by Silverman's rule of thumb, and the number that lead to the absolute highest tau score, and the final column shows how close, relatively, the tau score of the former setting is to the latter. Especially for the *Adult* and *Communities* datasets the relative score is high.

Dataset	histogram type	result	bins	τ	$\frac{\tau_{silverman}}{\tau_{max\ \tau}}$
Adult	equal-height	Silverman	44	0.971	0.990
		maximum τ	25	0.981	
	equal-width	Silverman	44	0.965	0.984
		maximum τ	27&29	0.981	
Communities	equal-height	Silverman	19	0.944	0.983
		maximum τ	11	0.960	
	equal-width	Silverman	19	0.936	0.965
		maximum τ	9	0.970	
Waterquality	equal-height	Silverman	20	0.807	0.893
		maximum τ	5	0.904	
	equal-width	Silverman	20	0.817	0.897
		maximum τ	7	0.910	

with respect to the absolute highest score is also the lowest of all three datasets. However, it is still only some ten percent away from this maximum, so in terms of quality, this is not a dramatic impact.

Another general trend that is visible in the correlation plots is that for both the EH and EW histograms, the suggested number of bins is roughly twice as big as the number of bins that lead to the absolute highest tau score. This high number is a direct consequence of the large number of examples in these datasets. A large n in Eq. 5 suggests that sufficient statistical support is present to afford a small kernel and small bins, and hence many bins to cover the complete domain of the target. Unfortunately, this heuristic does not take into account the downside of placing bin boundaries between adjacent values. For example, attribute *Age* of *Adult* has only 74 distinct values, such that a greater number of bins will always put every value in a separate bin, effectively rendering the attribute nominal[6]. Generally, a complicating factor for many datasets is the fact that the target attributes are not absolutely continuous (see Table 1), a feature that is assumed by Silverman's rule of thumb, and many algorithms dealing with continuous data.

Further note that having too many bins is also suboptimal from a computational point of view, as more bins need to be created and compared. But fortunately, these operations have very low complexity, see Section 4.2, and do not severely impact the search.

7. Discussion

This section reflects on the research questions and the experimental results presented in Section 6. A number of insights and suggestions for future work are offered that arose from the work presented above.

[6]The fact that higher numbers introduce empty bins plays no role here.

First point to address is the discrepancy between determining the bandwidth h for histogram and KDE experiments. In the histogram setting, bin boundaries are computed beforehand and used for each subgroup. Failing to do so would lead to incomparable density estimates, as the histograms of the subgroups and the data differ in number of bins or bin boundary locations. So, the histograms are constructed based on the dataset distribution, and may be suboptimal for the subgroups. In the KDE setting, a new bandwidth h is computed for each subgroup, as it provides the best density estimate. This setting is not plagued by the above-mentioned issue in the histogram approach, as the resulting density estimates are 'continuous'. This difference in methodology does not seem to influence the results too much, as the histogram versus KDE rank correlations are still high.

Q1 The main question addressed in this work is whether considering complete distributions allows for the discovery of interesting subgroups in data otherwise overlooked by traditional methods solely based on simple statistics. As results in Section 6.2 show, the proposed method does indeed value subgroups that are not rated high by the commonplace quality measure z-score. Moreover, results even make it strikingly clear that *none* of the results rated high by z-score, were deemed uninteresting by the squared Hellinger measure, and this holds true over increasing search depths. So, even as the squared Hellinger measure detects additional, previously uncovered, subgroups, no cost is incurred in this process in the form of subgroups missed by squared Hellinger, but recognised by z-score. This follows from the fact that there is general agreement between the measures about what subgroups are interesting with respect to mean shifts, but the new method additionally introduces the ability to detect distributional differences, like skewness and modality, in lieu of a difference in mean.

Q2 Beyond its application in the traditional Subgroup Discovery setting, multivariate probability density estimates enable the deployment of the new method within the Exceptional Model Mining framework. Section 6.3.1 offers corroborative results that the ability to recognise interesting subgroups with respect to centrality, as well as the possibility to value subgroups with other distributional differences, like multimodality, hold true also for higher-dimensional targets.

Two measures were considered, H^2 and L^2, and although both allow for the discovery of interesting subgroups, results were inconclusive with regard to the aim of understanding the different behaviours of these two measures, and to gauge to what extent they capture the same information. Results for the *Mammals* dataset suggest an (inverse) relation with respect to subgroup size, but no such correlation was found for the *Boston Housing* dataset. Experiments on these, and datasets not presented here, were unable to establish any clear relation between a subgroup's ranking and its size, for either method. Additional aspects that proved inconclusive predictors of both the behaviour of these individual measures, and their relation to each other, include subgroup to complement size ratios, their covariance, and subgroup or complement/data variance and variance ratios.

Finally, run time measurements do confirm that the analytical solution to density difference estimation offered by L^2 makes it orders of magnitude faster than the naive implementation employed in the H^2 solution.

Q3 For the purpose of measuring differences over complete distributions, two methods to estimate probability distributions were proposed: one based on histograms and one based on KDE. Section 6.4 indicates that the results produced by both methods generally agree with each other. In fact, overall, histograms perform (surprisingly) well, even

when using the number of bins suggested by Silverman's rule of thumb (*Q4* shows that this number is suboptimal with respect to correlation to the KDE ranking). However, robustness of histogram-based techniques is a known issue [11], and Figures 3 and 22 show histograms that, through a slightly different bin boundary placement, produce very different density estimates. So, it could be argued that the less complex histogram-based approaches are preferred, as long as one is attentive to subgroups incorrectly ranked. Of course, if complexity is not an issue, KDE is preferred, as its results require no additional scrutiny.

Q4 Section 6.5 addressed the effect of the pragmatic choice of using SilvermanâĂŹs rule of thumb to determine the number of bins for the histograms. Figure 23 shows that using the number of bins suggested by SilvermanâĂŹs rule of thumb does not lead to the optimal tau score in any of the experiments. The 'Silverman' setting achieved very good results for the *Adult* and *Communities* dataset, with a tau within 2% of the absolute highest value. For the *Waterquality* dataset, the result is about 10% from this absolute highest score, which is still reasonable.

A general observation is that the suggested number of bins is roughly twice the number of bins that lead to the absolute highest tau score. This is a somewhat unexpected result, as high bin counts result from a *small* bandwidth h, whereas Silverman is known to often *over-smooth*, something caused by a *large* h instead.

A number of alternative methods to determine the number of bins can be considered. First, alternative Silverman formulas exist that are capable of dealing with different distribution types [47], like non-normal or multimodal distributions. One such formula, $0.9 \cdot n^{-0.2} \cdot \min(\sigma, \mathrm{IQR}/1.34)$, uses the minimum of the standard deviation and the interquartile range (IQR). However, for the tested datasets, the alternative formula actually suggests a higher number of bins than the formula used in Eq. 5.

Another suggestion is to put a higher limit on the number of bins, such that on average bins should contain at least five percent of the data, for example. Putting a limit is also particularly relevant for datasets that have a low ratio of the target attribute cardinality with respect to dataset size (see Section 6.5).

A final suggestion is to use more complex *bandwidth selection strategies*. In their survey of the field of KDE, Jones et al. [49] make a distinction between first and second generation estimators. Among the first generation estimators are rule of thumb estimators, such as those of Silverman [47] and Scott [50]. More elaborate estimators are based on minimising the (asymptotic) mean integrated squared error (AMISE or MISE). Within these techniques, the deviation from the unknown true underlying distribution is initially guessed, and then optimised [47]. Second generation estimators include (biased [51]) cross-validation [52, 53] plug-in schemes, for example to improve the rate of convergence from the initial guess towards the true underlying density.

Although first generation estimators have their limitations, in the context of Subgroup Discovery, the main problem of second generation estimators is their high computational complexity. This is because a new PDF needs to be constructed for each and every subgroup tested, and in general also for each accompanying complement. As a typical experiment easily considers thousands of candidates, one generally would like to keep the estimation step itself relatively light. However, if this high computational demand is not prohibitive, using second generation (plug-in) estimators is a viable alternative.

Notwithstanding, overall, employing Silverman's rule of thumb to determine the number of bins was found to be adequate, both in terms of speed and result quality.

8. Conclusions

The main focus of this work was the introduction of an alternative method of measuring distributional differences in the context of Subgroup Discovery. Since popular quality measures for numeric targets were expected to be ill-suited for certain types of distribution, it proposes a new approach to recognise interesting deviations in subgroups. For the single target case, results show that a more complete evaluation of subgroup distributions does indeed counter a number of drawbacks associated with limited key statistics approaches. Additionally, the new measures do not only recognise interesting subgroups that were previously overlooked, but still adequately identify the subgroups judged interesting by traditional measures.

For the multivariate case, there is no existing work to compare to. Instead, two distance measures were compared to get an intuition of how they differ in terms of the results they produce. Although both measures identify subgroups in their own way, they both were demonstrated to identify interesting subgroups with multivariate, and potentially multimodal, deviations. Furthermore, as the L^2 measure was obtained as an analytical result, with a much lower computational complexity, it computes results orders of magnitude faster than H^2. It even performs swiftly in high-dimensional target settings. In short, the new evaluation strategy is a valuable addition to the Subgroup Discovery (and Exceptional Model Mining) framework.

References

[1] M. Atzmüller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.

[2] P. A. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

[3] W. Klösgen, "EXPLORA: A multipattern and multistrategy discovery assistant," in *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds.), pp. 249–271, Menlo Park, CA, USA: AAAI/MIT Press, 1996.

[4] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *PKDD 1997, Principles of Data Mining and Knowledge Discovery, European Symposium, Trondheim, Norway, 24–27 June, 1997, Proceedings* (H. J. Komorowski and J. M. Zytkow, eds.), vol. 1263 of *LNCS*, pp. 78–87, Springer, 1997.

[5] M. Atzmüller and F. Lemmerich, "Fast subgroup discovery for continuous target concepts," in *ISMIS 2009, International Symposium on Methodologies for Intelligent Systems, Prague, Czech Republic, 14–17 September, 2009, Proceedings* (J. Rauch, Z. W. Ras, P. Berka, and T. Elomaa, eds.), vol. 5722 of *LNCS*, (Berlin, Heidelberg, Germany), pp. 35–44, Springer, 2009.

[6] F. Lemmerich, M. Atzmüller, and F. Puppe, "Fast exhaustive subgroup discovery with numerical target concepts," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 711–762, 2016.

[7] B. F. I. Pieters, A. Knobbe, and S. Džeroski, "Subgroup discovery in ranked data, with an application to gene set enrichment," in *PL-10, Preference Learning Workshop at ECML PKDD 2010, Barcelona, Spain, 20–24 September, 2010*, 2010.

[8] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," in *ECML PKDD 2008, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Antwerp, Belgium, 15–19 September, 2008, Proceedings, Part II* (W. Daelemans, B. Goethals, and K. Morik, eds.), vol. 5212 of *LNCS*, (Berlin, Heidelberg, Germany), pp. 1–16, Springer, 2008.

[9] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[10] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[11] M. P. Wand and M. C. Jones, *Kernel Smoothing*. No. 60 in Monographs on Statistics and Applied Probability, Boca Raton, FL, USA: Chapman & Hall/CRC, 1994.

[12] A. M. Jorge, P. J. Azevedo, and F. Pereira, "Distribution rules with numeric attributes of interest," in *PKDD 2006, European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, 18–22 September, 2006, Proceedings* (J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), vol. 4213 of *LNCS*, pp. 247–258, Springer, 2006.

[13] G. Fasano and A. Franceschini, "A multidimensional version of the Kolmogorov–Smirnov test," *Monthly Notices of the Royal Astronomical Society*, vol. 225, no. 1, pp. 155–170, 1987.

[14] J. A. Peacock, "Two-dimensional goodness-of-fit testing in astronomy," *Monthly Notices of the Royal Astronomical Society*, vol. 202, no. 3, pp. 615–627, 1983.

[15] H. Grosskreutz, "Cascaded subgroups discovery with an application to regression," in *LeGo-08, From Local Patterns to Global Models Workshop at ECML PKDD 2008, Antwerp, Belgium, 15–19 September, 2008*, 2008.

[16] M. Boley, B. R. Goldsmith, L. M. Ghiringhelli, and J. Vreeken, "Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery," *Data Mining and Knowledge Discovery*, vol. 31, no. 5, pp. 1391–1418, 2017.

[17] F. Lemmerich, M. Becker, and F. Puppe, "Difference-based estimates for generalization-aware subgroup discovery," in *ECML PKDD 2013, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Prague, Czech Republic, 23–27 September, 2013, Proceedings, Part III* (H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, eds.), vol. 8190 of *LNCS*, (Berlin, Heidelberg, Germany), pp. 288–303, Springer, 2013.

[18] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *ICML 1995, International Conference on Machine Learning, Tahoe City, California, USA, 9–12 July, 1995, Proceedings* (A. Prieditis and S. J. Russell, eds.), (San Francisco, CA, USA), pp. 194–202, Morgan Kaufmann, 1995.

[19] M. Atzmüller and F. Lemmerich, "VIKAMINE – Open-Source subgroup discovery, pattern mining, and analytics," in Flach *et al.* [54], pp. 842–845.

[20] M. Meeng and A. Knobbe, "For real – A thorough look at numeric attributes in subgroup discovery," *Data Mining and Knowledge Discovery*, 2020.

[21] H. Grosskreutz and D. Paurat, "Fast and memory-efficient discovery of the top-k relevant subgroups in a reduced candidate space," in *ECML PKDD 2011, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Athens, Greece, 5–9 September, 2011, Proceedings, Part I* (D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, eds.), vol. 6911 of *LNCS*, (Berlin, Heidelberg, Germany), pp. 533–548, Springer, 2011.

[22] F. Lemmerich, M. Becker, and M. Atzmüller, "Generic pattern trees for exhaustive exceptional model mining," in Flach *et al.* [54], pp. 277–292.

[23] W. Duivesteijn, A. Feelders, and A. Knobbe, "Different slopes for different folks – Mining for exceptional regression models with Cook's distance," in *KDD 2012, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August, 2012, Proceedings* (Q. Yang, D. Agarwal, and J. Pei, eds.), pp. 868–876, ACM, 2012.

[24] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, May 16–18, 2000* (W. Chen, J. F. Naughton, and P. A. Bernstein, eds.), pp. 1–12, ACM, 2000.

[25] W. Duivesteijn, A. Knobbe, A. Feelders, and M. van Leeuwen, "Subgroup discovery meets Bayesian networks – An exceptional model mining approach," in *ICDM 2010, IEEE International Conference on Data Mining, Sydney, Australia, 14–17 December, 2010, Proceedings* (G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, eds.), (Los Alamitos, CA, USA), pp. 158–167, IEEE Computer Society, 2010.

[26] E. Galbrun and P. Miettinen, *Redescription Mining*. Springer Briefs in Computer Science, 2017.

[27] M. Meeng, W. Duivesteijn, and A. Knobbe, "ROCsearch – An ROC-guided search strategy for subgroup discovery," in *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24–26, 2014* (M. J. Zaki, Z. Obradovic, P. Tan, A. Banerjee, C. Kamath, and S. Parthasarathy, eds.), pp. 704–712, SIAM, 2014.

[28] M. van Leeuwen and A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 208–242, 2012.

[29] M. Meeng and A. Knobbe, "Flexible enrichment with Cortana – Software demo," in *Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning, Benelearn 2011, The Hague, The Netherlands, May 20, 2011* (P. van der Putten, C. Veenman, J. Vanschoren, M. Israel, and H. Blockeel, eds.),

pp. 117–119, 2011.

[30] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.

[31] P. B. Gibbons, Y. Matias, and V. Poosala, "Fast incremental maintenance of approximate histograms," in *VLDB 1997, Proceedings of the Twenty-third International Conference on Very Large Data Bases, Athens, Greece, August 25–29, 1997* (M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, eds.), pp. 466–475, Morgan Kaufmann, 1997.

[32] Y. E. Ioannidis, "The history of histograms (abridged)," in *VLDB 2003, International Conference on Very Large Data Bases, Berlin, Germany, 9–12 September, 2003, Proceedings* (J. C. Freytag, P. C. Lockemann, S. Abiteboul, M. J. Carey, P. G. Selinger, and A. Heuer, eds.), Morgan Kaufmann, 2003.

[33] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *IJCAI 1993, International Joint Conference on Artificial Intelligence, Chambéry, France, August 28 – September 3, 1993* (R. Bajcsy, ed.), pp. 1022–1029, Morgan Kaufmann, 1993.

[34] P. Kontkanen and P. Myllymäki, "MDL histogram density estimation," in *AISTATS 2007, International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico, 21–24 March, 2007, Proceedings* (M. Meila and X. Shen, eds.), vol. 2 of *JMLR Proceedings*, pp. 219–226, JMLR.org, 2007.

[35] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

[36] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[37] S. J. Sheather, "Density estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, 2004.

[38] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Verlag, 2006.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[40] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[41] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, no. 1, pp. 41–54, 1994.

[42] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning, MIT Press, 2006.

[43] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[44] S. Džeroski, D. Dešmsar, and J. Grbović, "Predicting chemical parameters of river water quality from bioindicator data," *Applied Intelligence*, vol. 13, no. 1, pp. 7–17, 2000.

[45] G. C. Garriga, H. Heikinheimo, and J. K. Seppänen, "Cross-mining binary and numerical attributes," in *ICDM 2007, IEEE International Conference on Data Mining, Omaha, Nebraska, USA, 28–31 October, 2007, Proceedings* (N. Ramakrishnan, O. R. Zaïane, Y. Shi, C. W. Clifton, and X. Wu, eds.), (Los Alamitos, CA, USA), pp. 481–486, IEEE Computer Society, Oct. 2007.

[46] A. J. Mitchell-Jones, G. Amori, W. Bogdanowicz, B. Krystufek, P. J. H. Reijnders, F. Spitzenberger, M. Stubbe, J. B. M. Thissen, V. Vohralik, and J. Zima, *The Atlas of European Mammals*, vol. 3. Academic Press London, 1999.

[47] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[48] L. Denby and C. Mallows, "Variations on the histogram," *Journal of Computational and Graphical Statistics*, vol. 18, no. 1, pp. 21–31, 2009.

[49] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 401–407, 1996.

[50] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[51] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society*, vol. 53, no. 3, pp. 683–690, 1991.

[52] A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, no. 2, pp. 353–360, 1984.

[53] P. Hall, J. S. Marron, and B. U. Park, "Smoothed cross-validation," *Probability Theory and Related Fields*, vol. 92, no. 1, pp. 1–20, 1992.

[54] P. A. Flach, T. De Bie, and N. Cristianini, eds., *ECML PKDD 2012, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Bristol, UK, 24–28 September, 2012, Proceedings (II)*, vol. 7524 of *LNCS*, (Berlin, Heidelberg, Germany), Springer, 2012.