BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 62, No. 2, 2014
DOI: 10.2478/bpasts-2014-0029

Dedicated to Professor J. Klamka
on the occasion of his 70th birthday

Query-condition-aware V-optimal histogram in range query selectivity estimation

D.R. AUGUSTYN*

Institute of Computer Science, Silesian University of Technology, 16 Akademicka St., 44-100 Gliwice, Poland

Abstract. Obtaining the optimal query execution plan requires a selectivity estimation. The selectivity value allows to predict the size of a query result. This lets choose the best method of query execution. There are many selectivity obtaining methods that are based on different types of estimators of attribute values distribution (commonly they are based on histograms). The adaptive method, proposed in this paper, uses either attribute values distribution or range query condition boundaries one. The new type of histogram – the Query-Conditional-Aware V-optimal one (QCA-V-optimal) – is proposed as a non-parametric estimator of a probability density function of attribute values distribution. This histogram also takes into account information about already processed queries. This information is represented by the 1-dimensional Query Condition Distribution histogram (HQCD) which is an estimator of the include function P_I which is also introduced in this paper. P_I describes so-called regions of user interest, i.e. it shows how often regions of attribute values domain were used by processed queries. Advantages of the proposed method based on QCA-V-optimal are presented. Conducted experiments reveal small values of a mean relative selectivity estimation error comparing to the error values obtained by methods based on the relevant classical V-optimal histogram and Equi-height one.

Key words: query selectivity estimation, attribute values distribution, Query-Condition-Aware V-optimal histogram.

1. Introduction

Query execution in Database Management System (DBMS) consists of two phases – a prepare (parse) phase and an execution one. One of the most important activity in the prepare phase is called a query optimization process. During this, the best method of a query execution (among many other possible methods) is chosen by a cost-based query optimizer (CBQO). CBQO obtains so-called an execution plan – the method of a query execution which satisfies the lowest cost criterion. A query cost may be measured using many factors but the most important one is the size of the data that should be retrieved from a database. This is a reason why a selectivity parameter is introduced.

Selectivity for a single-table query with a selection condition based on one attribute is the number of rows satisfying the selection condition divided by the number of all rows in a table. The selectivity is also a probability of drawing a row satisfying the selection condition from the set of all table rows.

The selectivity parameter for a single-table range query Q with a selection condition based on a one attribute X with continuous domain may be obtained from:

$$sel(Q(a < X < b) = \int_a^b f_x(x)dx, \tag{1}$$

where $f_x(x)$ is a probability density function (PDF) describing X distribution. So we can see that some estimator of PDF is needed for selectivity calculation. Since many years, histograms have been used in DBMSs as non-parametric estimators of PDF [1].

Most of work related to the selectivity estimation problem concentrates on obtaining selectivity for queries with a complex selection condition based on many attributes. This is the problem of selectivity calculation for so-called multidimensional query condition. Thus a multidimensional estimator of a multivariate PDF is required. Here, the most important problem is finding a space-efficient multidimensional distribution representation. Since years, there are known many approaches to obtaining a small-sized representation of multidimensional PDF estimator like those ones: PHASED [2, 3], MHIST [2, 3], GENHIST [3], multidimensional kernel estimator [3, 4], kernel spline [5], Bayesian Network [6], Discrete Cosine Transform [7], Cosine Series [8], Discrete Wavelets Transform [9], Self-tuning histogram [10–12].

Some researches concentrate on a problem of prediction of representation of an attribute values distribution. This is important when a current representation cannot be directly built on values from a database (e.g. because high Input-Output workload). A prediction method requires to have an information about previous representations (created in previous moments of execution of update statistics operations). This requires also a mathematical model of time evolution of attribute values distribution. It may be based on tracking previous values of non-central moments [18] or previous locations of quantiles [19]. A dynamical system (especially fluid-flow-based one) or an artificial neural network (with a different architecture) may be used as a model of evolution of the mentioned above parameters of distribution. The approaches that

*e-mail: draugustyn@polsl.pl

utilize dynamical systems or neural networks are also applied in other fields of computer science, e.g. [17, 20, 21].

The additional direction of a research on the selectivity estimation refers to taking into account not only an attribute values distribution but also a distribution of selection condition boundaries of already executed queries. Somewhat the adaptive method based on self-tuning histograms could be also qualified to such approach. Self-tuning histograms use the size of results of already processed queries to refine themselves.

Another considered adaptive approach is HQCA-based method (Query-Condition-Aware histogram-based method) [13]. Here we use information about conditions of already executed queries (not about the sizes of results of already executed queries like in [10, 11]). In this method we collect information either about attribute value distribution or query condition boundaries distribution (using Equi-width HQCD histogram (Query-Condition-Distribution histogram)). The method of selectivity estimation may be used only for 1-dimensional queries. 1-dimensional range queries are described by a 2-dimesional distribution of query boundaries. Because of the mentioned space-saving reason this 2-dimesional distribution of query boundaries is represented here by an approximate 1-dimesional description which is estimated by HQCD. Then HQCD is used to divide X domain into N_{qcd} intervals. This is quantile division of X domain. It is obvious that intervals are narrow in regions where HQCD values are high. Those are so-called regions of high user interest. In each interval an Equi-width subhistogram is created. The number of buckets in a subhistogram equals N_{eqb} and it is the same for all intervals. All buckets define the final HQCA histogram. The number of HQCA buckets equals $B = N_{qcd} \cdot N_{eqb}$. HQCA is neither an Equi-depth histogram nor an Equi-width one. Sizes of HQCA buckets are small in regions of high user interest. So HQCA resolution is high in those X regions where query condition intervals extensively overlap them. Some disadvantage of this method results from the problem of B factorization – the method cannot work for any B values. Those values of B which have many factorizations should be used. The new proposed approach based on Query-Condition-Aware V-optimal histogram (QCA-V-optimal) overcomes this problem.

In the method described in this paper we define a QCA-V-optimal histogram. The proposed method is based on V-optimal approach [14] and it extends V-Optimal histogram by introducing an information about query condition bounds distribution. In this method, the mentioned HQCD histogram is used for modifying the error metric formula (which is used in the classical Voptimal creation algorithm). The dynamical programming method proposed in V-optimal histogram creation is also used in QCA-V-optimal histogram creation.

The main contributions of the paper are:

- introducing the formal definition of P_I – the include function – which approximately space-efficiently describes a 2-dimensional query condition boundaries distribution (the mentioned HQCD is a non-parametric estimator of P_I) (Sec. 3),
- introducing QCA-V-optimal histogram type (more flexible than HQCA [13]) representing either attribute values distribution or query condition boundaries one (Sec. 4),
- experimental results presenting advantages of the selectivity estimation method based on QCA-V-optimal histogram (i.e. the smallest average mean relative selectivity estimation error comparing to the error values obtained using V-optimal histogram and Equi-depth one for a given constant size of the distribution representation) (Sec. 7).

The paper is organized as follows. Section 2 shortly presents the known-method of describing an attribute values distribution using a V-optimal histogram. In Sec. 3 we introduce the proposal of including function and its estimator – HQCD histogram. They approximately describe boundaries of range query conditions. In Sec. 3 we show how to build a HQCD histogram using boundaries values from range conditions of already processed queries. Section 4 describes a new type of histogram i.e. the QCA-V-optimal one which is based on either a distribution of table attribute values or a range query conditions distribution. Steps of the algorithm of creating a QCA-V-optimal histogram are presented in Sec. 5. Section 6 shows exemplary concrete distributions of attribute values (Subsec. 6.1) and query range boundaries values (Subecs. 6.2–6.5). Those distributions are used in Sec. 7 for experimental verifying an accuracy of query selectivity estimations based on QCA-V-optimal histogram, V-optimal one, and Equi-depth one. Subsections 7.1–7.4 show either result QCA-V-optimal histograms or result values of mean relative selectivity estimation error for the boundaries values distributions assumed in Subsecs. 6.2–6.5. Subsection 7.5 gives a synthetic view on all obtained experimental results.

2. V-optimal histogram – non-parametric estimator of attribute value distribution

V-optimal histogram [14] is a well-known approach to estimate a probability distribution and obtain an approximate selectivity value.

Let us assume the following notations and definitions (based on [14]):

X – attribute of relation R (with real or integer domain),

V – sequence of unique X values that exist in relation R, $V = (v_i)_{i=1}^{N}$ where $\underset{j>i}{\forall} v_i < v_j$,

$f(v)$ – frequency of v occurrence i.e. the number of tuples $t \in R$ where $t.X = v$,

F – frequency vector, $F = (v_i)_{i=1}^{N}$ where $f_i = f(v_i)$,

B – number of histogram buckets where $B < N$,

b_j, e_j – indexes of endpoints (begin, end) of the j-th bucket interval where $j = 1...B \wedge b_j \leq e_j \wedge b_j,\ e_j \in \{1...N\}$,

h_j – frequency approximation in the j-th bucket for $j = 1...B$:

$$h_j = \text{Avg}(b_j, e_j) = \frac{\sum\limits_{b_j \leq m \leq e_j} f_m}{e_j - b_j + 1}, \tag{2}$$

where $\text{Avg}(k, l)$ for $k, l = 1...N$ and $k \leq l$ is an average frequency of $f_k...f_l$.

SSE (k, l) – sum square error of frequencies f_m for $1 \leq k \leq m \leq l \leq N$:

$$\text{SSE}(k,l) = \sum_{k \leq m \leq l} (f_m - \text{Avg}(k,l))^2. \quad (3)$$

SSE (b_j, e_j) may be considered as a metric of frequency approximation error in the j-th bucket ($j = 1...B$).

SSE satisfies the rule:

$$\text{SSE}(k,l) \geq \text{SSE}(k,m) + \text{SSE}(m,l) \quad (4)$$

for $1 \leq k \leq m \leq l \leq N$.

Let us denote OSSE(m, k) – the optimal sum square error – as a minimum SSE for a subsequence of F: f_1, ..., f_m ($m \leq N$) using k buckets ($k \leq B$).

Because of the property ([14]):

$$\text{OSSE}(m,k) = \min_{1 \leq j \leq m} (\text{OSSE}(j, k-1) + \text{SSE}(j+1, m)), \quad (5)$$

the Bellman's principle and concept of dynamical programming may be used to obtain the optimal solution – just OSSE (N, B) should be determined. The problem of finding the optimal solution for k buckets of histogram may be reduced to finding the optimal solutions of subproblem for $k-1$ buckets.

Finally, obtaining OSSE(N, B) is equivalent to finding V-optimal histogram with B buckets for a given F vector. Calculating OSSE(N, B) allows to find required SSE-optimal set of endpoints (b_j, e_j) (where $j = 1...B$).

The presented-above algorithm and its improved faster version but also the approximate-optimal one were considered in [14].

We may also assume another interpretation of V which is useful for X with continuous domain. We may approximate X distribution using an Equi-width histogram. Then centers of intervals of the Equi-width histogram can be treated as elements of V. Values of the Equi-width histogram can be treated as elements of F.

3. Including function – description of range query boundaries distribution

3.1. Including function definition. Let us consider a space [0, 1]^2 as a domain of $a \times b$. This simplifies the problem description but it does not lessen generality of conclusions.

Let us introduce a functional $P_I(y, f)$ which approximately describes a range query boundaries distribution defined by probability density function $f(a, b)$. For a concrete $f(a, b)$ the P_I becomes a function of y, where y belongs to [0, 1]. P_I is probability that randomly chosen y value is included in any interval [a, b] defined by some range query (i.e. $0 \leq a \leq y$ and $y \leq b \leq 1$). P_I will be called *include function*.

The definition of P_I for all query bounds can be formulated as:

$$P_I(y,f) \stackrel{def}{=} \frac{1}{\text{number of pairs}} \sum_{\text{all possible pairs } (a,b)} P(a \leq y \leq b). \quad (6)$$

For a continuous joint distribution of a and b the P_I may be obtained as follows:

$$P_I(y,f) = \iint_{\text{Dom}(a,y) \times \text{Dom}(b,y)} f(a,b)dadb = \int_0^y \int_y^1 f(a,b)dbda, \quad (7)$$

where Dom(a, y) is a range of variation of a for a given y value, and Dom(b, y) is a range of variation of b for a given y value. The method of calculating value P_I as a definite double integral over some Dom(a, y) $\times$ Dom(b, y) region is illustrated in Fig. 1.

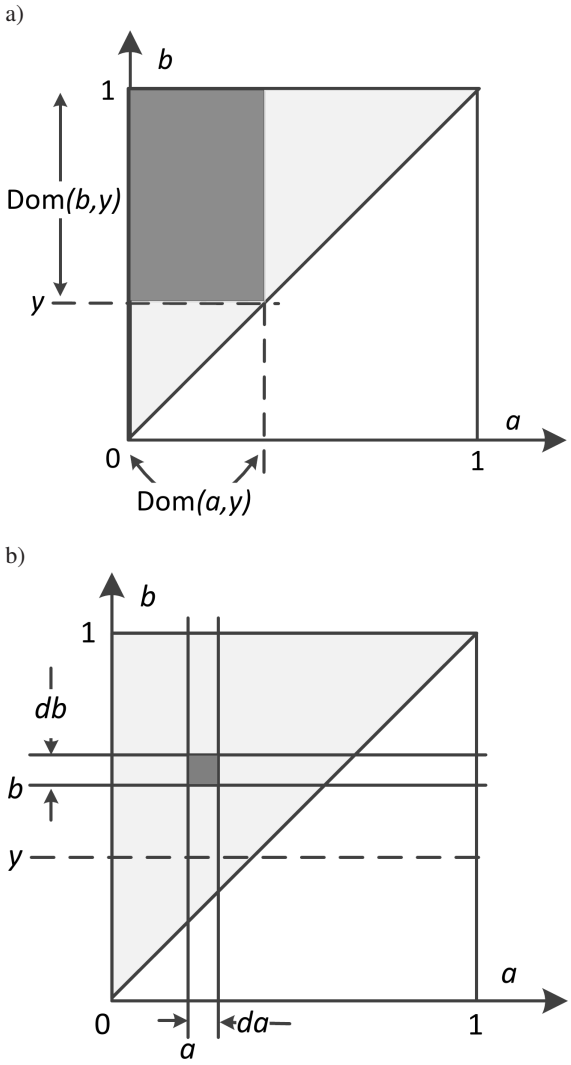

Fig. 1. a) domains of a and b values for given y value, b) differentials of a and b in integration of f

Univariate P_I is a lossy representation of bivariate f.

Because P_I (y, f) is based on integral operations it satisfies the linearity property, i.e.:

$$\begin{aligned} P_I(y, f_1 + f_2) &= P_I(y, f_1) + P_I(y, f_2) \wedge P_I(y, \alpha f) \\ &= \alpha P_I(y, f), \end{aligned} \quad (8)$$

where $\alpha = \text{const}$.

To show an example of some concrete P_I let us consider a simple 2-dimensional uniform distribution defined by $f_{2D-uniform}$ density function as follows:

$$f_{2D-uniform}(a,b)=\begin{cases} 2 \text{ for } 0 \le a \le 1 \wedge 0 \le b \le 1 \wedge b \le a \\ 0 \qquad \text{otherwise} \end{cases}. \tag{9}$$

It is shown in Fig. 2a.

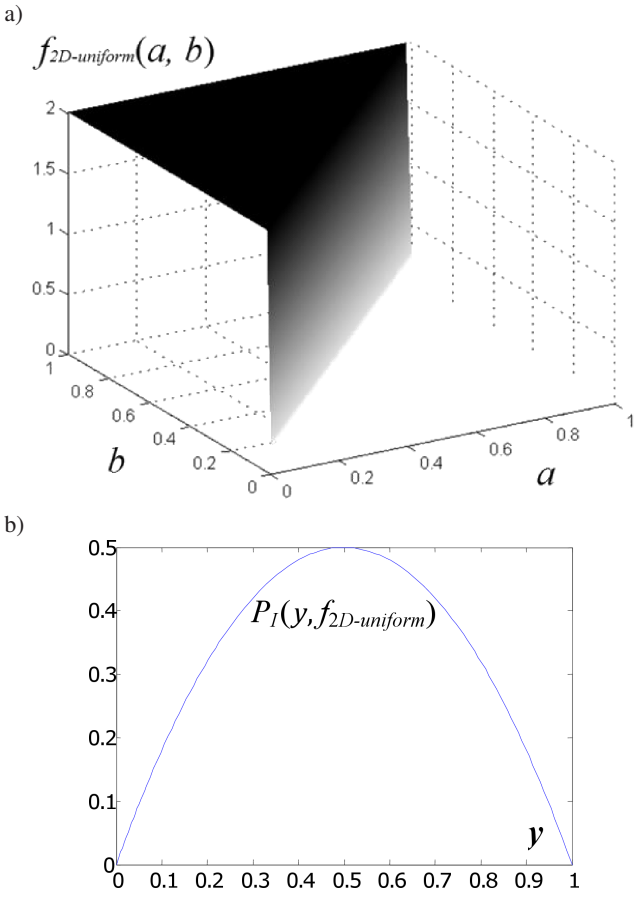

Fig. 2. a) bivariate probability density function $f_{2D-uniform}$, b) corresponding including function $P_I(y, f_{2Duniform})$

Using Eq. (7) and (9) we can find the include function for $f_{2D-uniform}$ as follows:

$$P_I(y, f_{2D-uniform}) = \int_0^y \int_y^1 2dbda = 2y(1-y). \tag{10}$$

$P_I(y, f_{2D-uniform})$ is shown in Fig. 2b.

3.2. HQCD – histogram estimator of including function. A non-parametric estimator of including function is called HQCD – histogram of a query condition boundaries distribution. HQCD histogram was proposed in HQCA-based approach in [13]. HQCD is an Equi-width histogram.

Values of HQCD buckets of may be updated during online query processing (i.e. during a prepare phase). If a query condition interval [a, b] overlaps some buckets of HQCD than values in those buckets are incremented (When a bucket is overlapped by the interval [a, b] in more than 50% of its length, a corresponding bucket value is incremented). The process of HQCD updating is presented in Fig. 3.

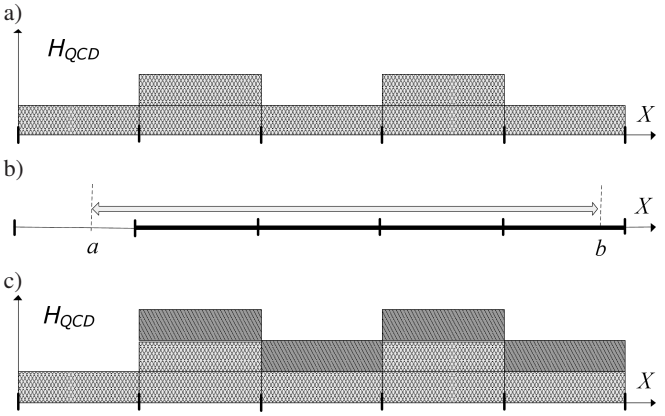

Fig. 3. Technique of a HQCD building: a) initial state of the HQCD, b) some interval of a query condition, c) HQCD state after taking into account the interval (source: [13] Fig. 4)

The last activity in HQCD creation process is done after data about range condition boundaries of processed queries is gathered. Every number from a HQCD bucket is divided by a total number of queries that modified any HQCD bucket.

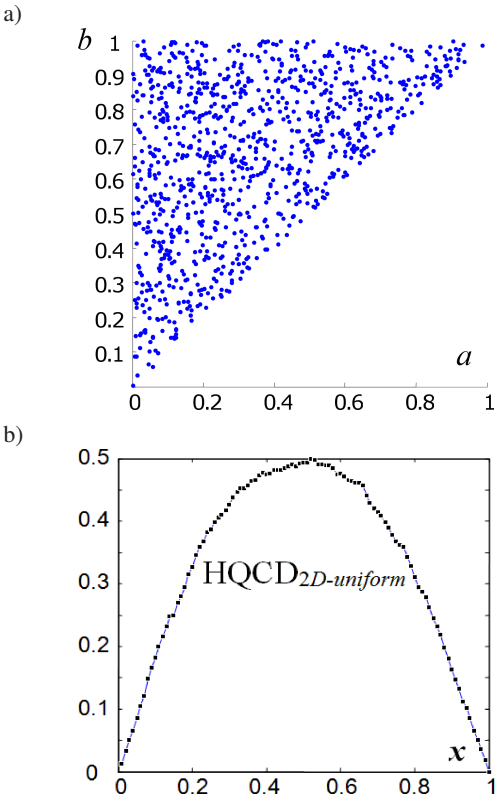

Fig. 4. a) sample set of (*a, b*) pairs based on 2*D-uniform* distribution, b) $HQCD_{2D-uniform}$ histogram estimating the include function for 2*D-uniform* distribution (the histogram based on the samples shown in Fig. 4a)

Figure 4a presents a sample data set which satisfies 2*D-uniform* distribution. The size of the sample set is $N_{qb} = 1000$. Figure 4b presents HQCD histogram built on this sample set. The histogram is an non-parametric estimator of $P_I(y, f_{2D-uniform})$ (shown if Fig. 2b). The number of $\text{HQCD}_{2D-uniform}$ buckets equals $N_q = 100$.

4. Query-Condition-Aware-V-optimal histogram – estimator based on either attribute value distribution or range query condition boundaries one

A new type of histogram – the Query-Condition-Aware-V-optimal histogram – is proposed in this section. This histogram is based on the V-optimal approach which was described in Sec. 2. However in QCA-V-optimal histogram either an attribute value distribution or a query boundaries distribution are taken into account. Include function is used as an approximate representation of a 2-dimensional distribution of query boundaries. The general method of histogram constructing is the same like the one described in Sec. 2 but the error metric is different than this given by Eq. (3).

A sum square error in a histogram interval is modified by a sum of include function values for frequencies belonging to this interval. This new error metric is denoted by SSEW (Weighted Sum Square Error) and defined by:

$$\begin{aligned} \text{SSEW}(k,l) &= \text{SSE}(k,l) \cdot \sum_{k \le m \le l} \text{HQCD}(m) \\ &= \sum_{k \le m \le l} (f_m - \text{Avg}(k,l))^2 \cdot \sum_{k \le m \le l} \text{HQCD}(m). \end{aligned} \tag{11}$$

HQCD is estimator of P_I function (according to the consideration in Sec. 2). HQCD(m) in Eq. (11) is the approximate value of include function for the m-th element of F.

Usage of $\sum_{k \le m \le l} \text{HQCD}(m)$ factor in Eq. (11) affects that even some intervals with high value of SSE may have small value of SSEW. Intervals with small values of P_I (i.e. intervals rarely used by queries) will have rather small values of SSEW. The method of QCA-V-optimal histogram constructing implicitly divides intervals with high values of SSEW. Such regions of X domain (with a high SSEW) are represented by a high resolution regions of QCA-V-optimal histogram.

QCA-V-optimal histogram satisfies the compromise between the criteria based on: including of X values into query condition intervals (described by HQCD) and a variance of X (described by differences between adjacent elements of F, i.e. described by SSE values).

5. Procedure of Query-Condition-Aware V-optimal histogram creating

Process of QCA-V-optimal histogram constructing consists of the following activities:

- turning on the process of gathering information about range query condition boundaries, i.e. starting HQCD buckets updating,
- turning off the HQCD updating process after some period of time,
- gathering information about attribute values distribution by creating F vector,
- creating QCA-V-optimal histogram based on either F vector or HQCD histogram (after that HQCD and F representations may be removed).

After finishing the described-above process we can turn on (enable to CBQO) the selectivity estimation method based on the newly created QCA-V-optimal histogram.

6. Validation data sets

This section describes sample distributions and data sets used for validation of the proposed selectivity estimation method. We use some synthetic data sets created by pseudorandom generators. Probability distributions use in experiments are described in details below. For 1-dimensional attribute values distribution (i.e. X distribution) we use Gaussian clusters (Subsec. 6.1). 2-dimensional range query condition boundaries distributions (i.e. distribution of pairs (a, b)) are based on:

- Narrow Interval preferred distribution (combined distribution: normal distribution of a and truncated exponential one of b) (Subsec. 6.2)
- Superposition of Gaussian clusters with one or two peaks (Subsec. 6.3 and 6. 4)
- Individual Uniform distribution (distribution of a is uniform; conditional distribution of b for given a value is also uniform) (Subsec. 6.5).

6.1. Sample attribute value distribution. X values distribution – a superposition of two normal distributions – is given by the probability density function:

$$\begin{aligned} f_x(x) = 0.5 \ &\text{PDF}(\text{N}(0.5, 0.06)) \\ + 0.5 \ &\text{PDF}(\text{N}(0.6, 0.002)) \end{aligned} \tag{12}$$

and it is shown in Fig. 5a (bold line). Figure 5b shows F vector obtained from a sample set values o X variable i.e. $f(v_i)$. The size of F vector equals $N = 100$. $N_x = 1000$ is a sample size used for F vector creation.

Basing on the method described in the Sec. 2, the V elements are grouped into buckets shown in Fig. 6. The number of buckets B equals 20. Vertical dashed lines in Fig. 6 show boundaries of bucket intervals.

Using the division shown in Fig. 6, the relevant V-optimal histogram was created and it is shown in Fig. 7. A representation of the histogram requires to allocate 41 numbers, i.e. $B+1$ boundaries of buckets and B histogram values (one for each bucket).

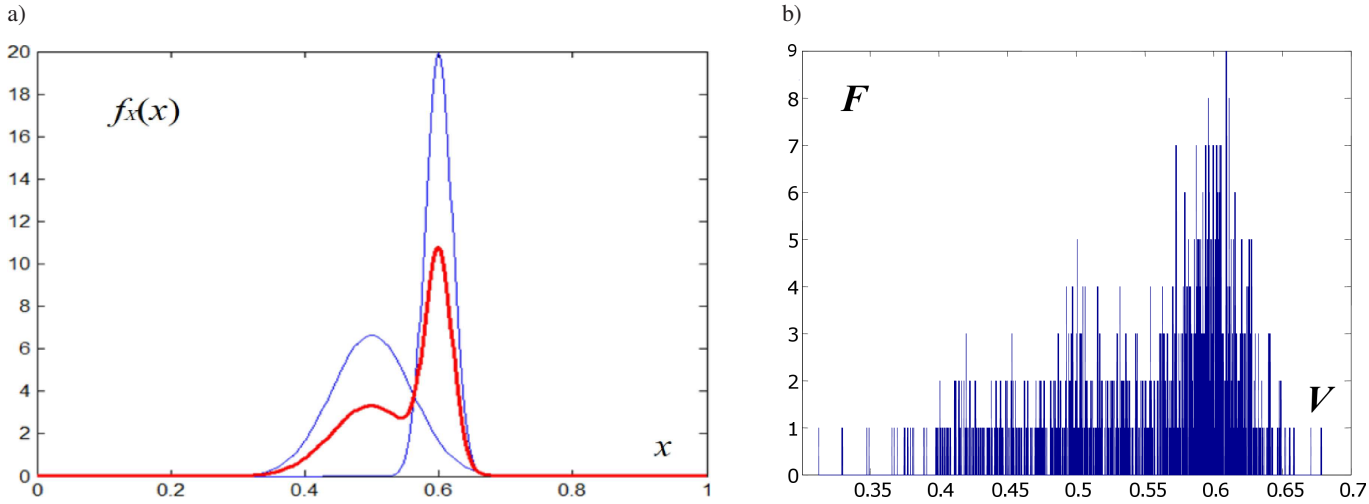

Fig. 5. a) probability density function of X variable distribution based on two 1D Gaussian clusters, b) $f(v_i)$ for $i = 1...100$ – the frequency vector built on the sample of X variable [13]

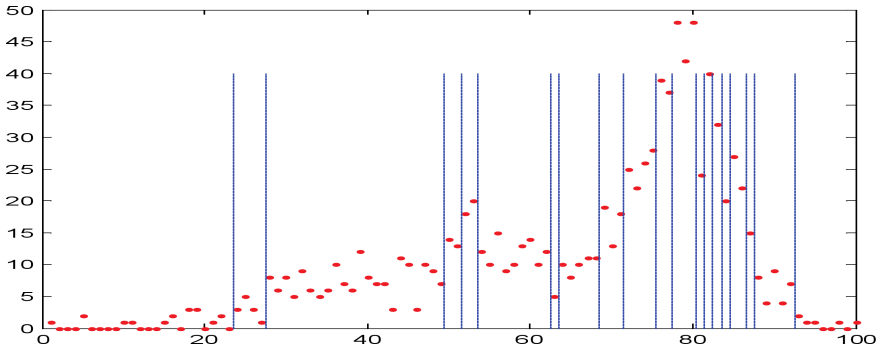

Fig. 6. Error optimal X domain division using SSE – the classical error metric given by Eq. (3). V elements (shown in Fig. 5b) grouped into buckets of V-optimal histogram

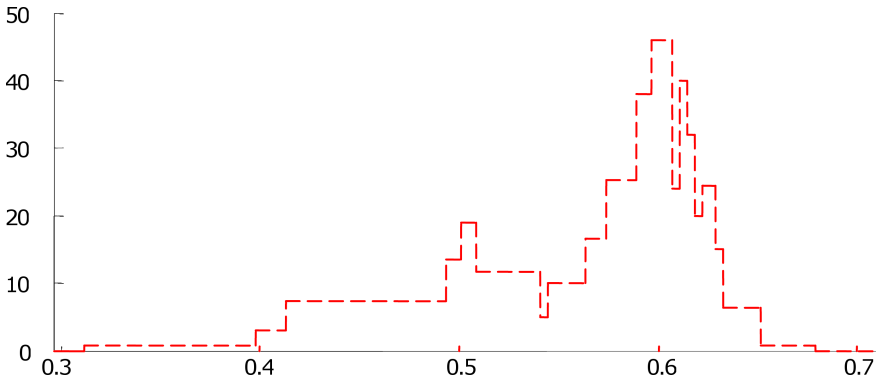

Fig. 7. Classical V-optimal histogram shown along the X distribution support (the number of bucket equals 20)

In our work we also consider a usage of an Equi-height histogram type (also called Equi-depth) for describing X distribution. This enables to compare the proposed selectivity estimation method to the method based on histogram that are commonly used in DBMSs (e.g. Oracle DBMS, MS SQL Server). Figure 8 shows the Equi-height histogram for X attribute based on the data set described by F vector. The number of buckets of this Equi-height histogram equals 40. This results from the general assumption that the size of a space-allocation for an Equi-height histogram and the size for a V-optimal one are the same.

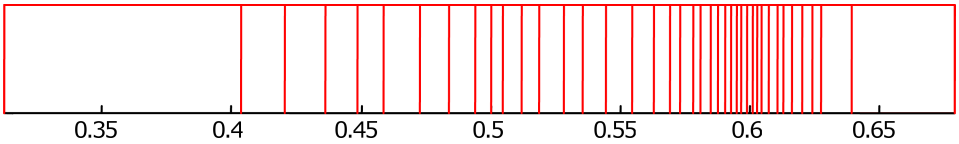

Fig. 8. Classical Equi-height histogram shown along the X distribution support (the number of bucket equals 40)

6.2. NI query boundaries distribution – Narrow Interval preferred distribution. (a, b) pairs of query condition endpoints (satisfying *NI* probability distribution) are generated as follows:

- a variable satisfies normal distribution N(0.6, 0.05),
- b value for a given a value can be obtained from: $b = (1 - a)z + a$,

where z is a length of a generated query interval.

z is described by a truncated exponential distribution. To generate a z value an additional variable z' will be introduced. A z' value is obtained from a pseudorandom generator which is based on the exponential distribution described by the probability density function $-f(z') = \lambda e^{-\lambda z'} 1(z')$ where $E(z') = \dfrac{1}{\lambda} = 0.1$. If z' is less or equal $z_{\max} = 1 - a$ then z' becomes z else a new z' value is generated etc. Thus z is described by the conditional probability density function as follows:

$$f(z|\,a) = f(z|\,z_{\max} = 1 - a) = \frac{\lambda e^{-\lambda z}}{1 - e^{-\lambda z_{\max}}} \{1(z) - 1(z - z_{\max})\} \tag{13}$$

and it is shown in Fig. 9b.

Mean value of z can be obtained from:

$$E(z) = \int_0^{z_{\max}} f(z)dz = \frac{1}{\lambda} \frac{(1 - (1 + \lambda z_{\max})e^{-\lambda z_{\max}})}{(1 - e^{-\lambda z_{\max}})}. \tag{14}$$

Because of the exponential distribution property (positive skewness property), small intervals are preferred. This means that generated query intervals z are rather smaller than $E(z|a)$. For example in Fig. 9b for $a = 0.7$ (i.e. $z_{\max} = 0.3$) we can see that z values are rather less than value $E(z|a) \approx 0.084$ (which is obtained from Eq. (14)) because $P(Z \leq E(z|_{a=0.7}) \approx 0.6 > P(Z > E(z|_{a=0.7}) \approx 0.4$. Thus results the name of the joint a and b distribution (this is a reason why "narrow interval" appears in the name of distribution).

Figure 9a shows a smoothed histogram describing *NI* query boundaries distribution. This 2-dimensional histogram was built on the sample set of (a, b) pairs presented in Fig. 10a. The size of the sample set is $N_{qb} = 1000$.

HQCD$_{NI}$ histogram based on the sample set (shown in Fig. 10a) is presented in Fig. 10b. The number of HQCD$_{NI}$ buckets equals $N_q = 100$. We can see that about 45% of X domain is out of HQCD support (i.e. only X values form 0.45 to 1 are included in ranges of queries described by *NI* probability distribution).

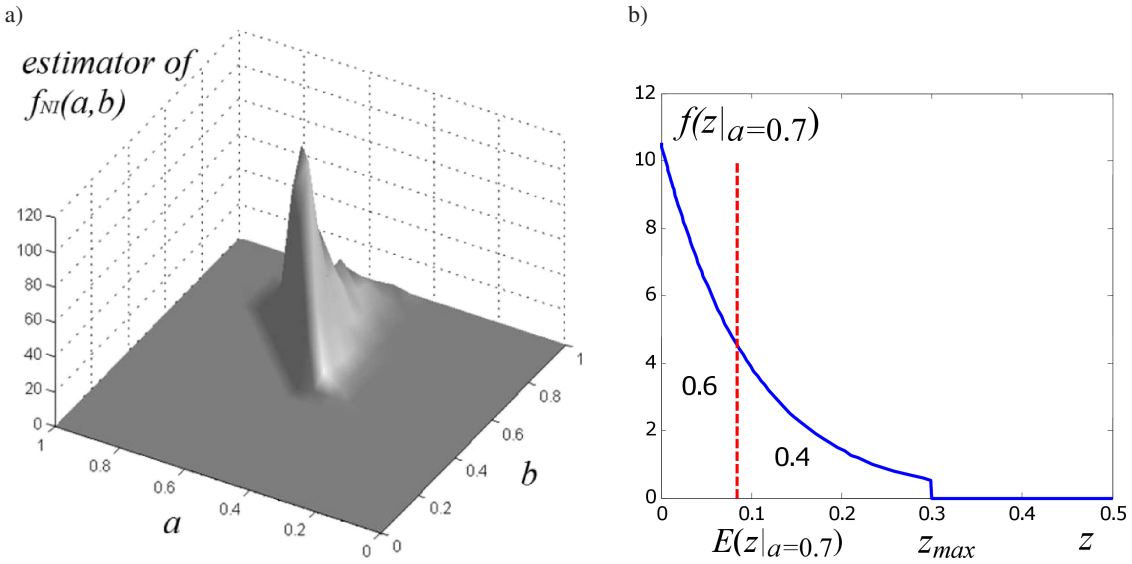

Fig. 9. a) smoothed histogram describing *NI* distribution based on the sample set of (a, b) pairs (shown in Fig. 10a), b) $f(z|a)$ – probability density function of distribution of length query interval $z = b - a$ for $a = 0.7$ (i.e. $z_{\max} = 0.3$)

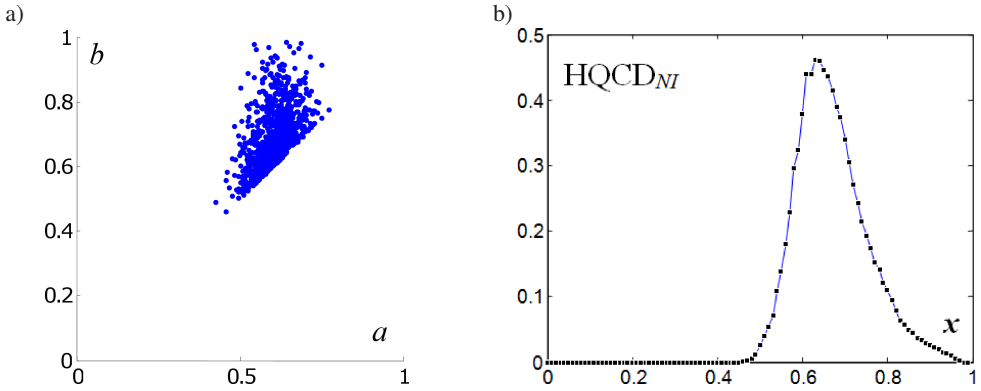

Fig. 10. a) sample set based on *NI* distribution, b) HQCD$_{NI}$ histogram estimating the include function for *NI* distribution (the histogram based on the samples shown in Fig. 10a)

6.3. 1GC query boundaries distribution – 1 Gaussian Cluster distribution. The probability density function of $1GC$ distribution of query boundaries – $f_{1GC}(a,b)$ is described by:

$$f_{1GC}(a,b) = \text{PDF}(\text{N}(m,\Sigma)) \tag{15}$$

with mean vector and covariance matrix:

$$m = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 0.0004 & -0.0001 \\ -0.0001 & 0.0001 \end{bmatrix} \tag{16}$$

is shown in Fig. 11a.

The sample set of query boundaries is presented in Fig. 11b and 12a. The size of the set is $N_{qb} = 1000$. The set was generated according to the distribution described by $f_{1GC}(a,b)$. A HQCD was created using this set. The number of intervals of HQCD equals $N_b = 100$. The HQCD$_{1GC}$ estimating the $P_I(y, f_{1GC})$ is shown in Fig. 12b.

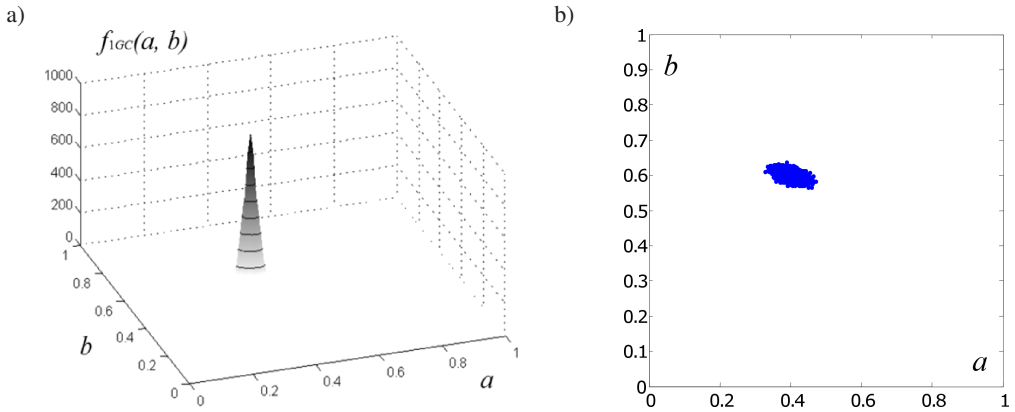

Fig. 11. a) $f_{1GC}(a,b)$ density function based on 1 Gaussian cluster, b) sample set of query boundaries generated using $f_{1GC}(a,b)$

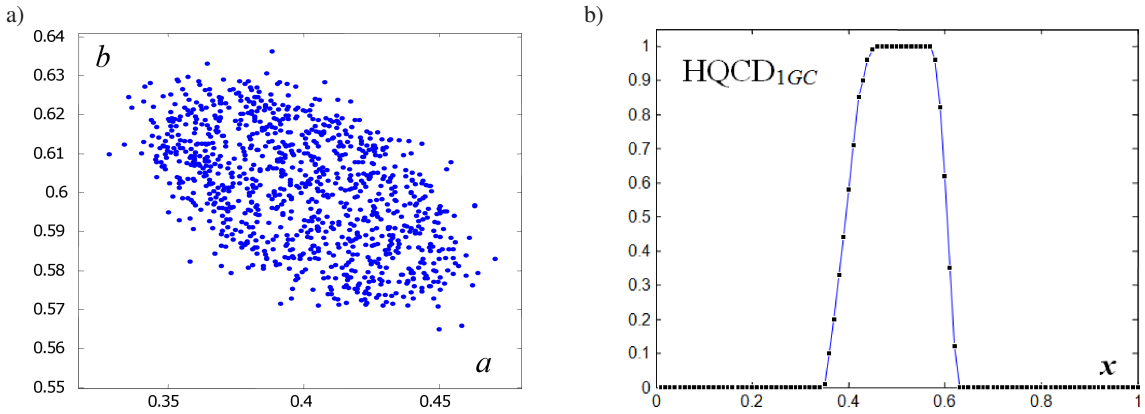

Fig. 12. a) zoomed view of the sample set of query boundaries (shown in Fig. 11b), b) HQCD$_{1GC}$ histogram estimating the include function $P_I(y, f_{1GC})$ based on the samples (shown in Fig. 11b and 12a)

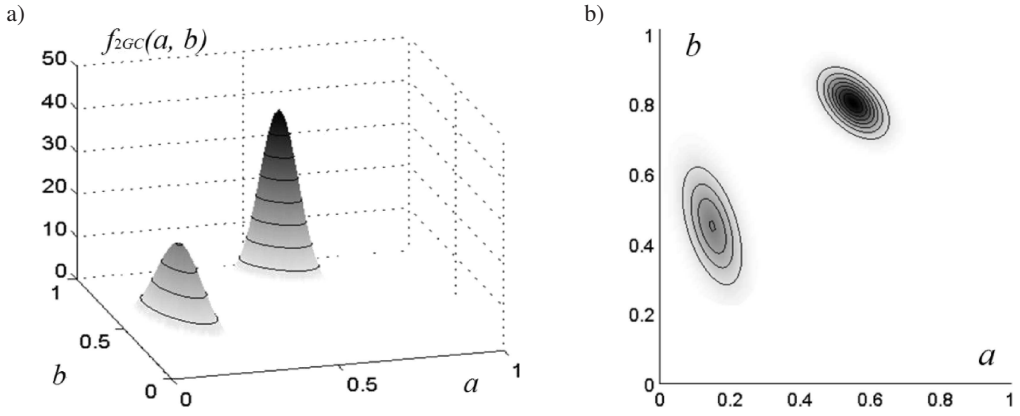

Fig. 13. a) $f_{2GC}(a, b)$ density function based on 2 Gaussian clusters, b) contour graph for $f_{2GC}(a, b)$

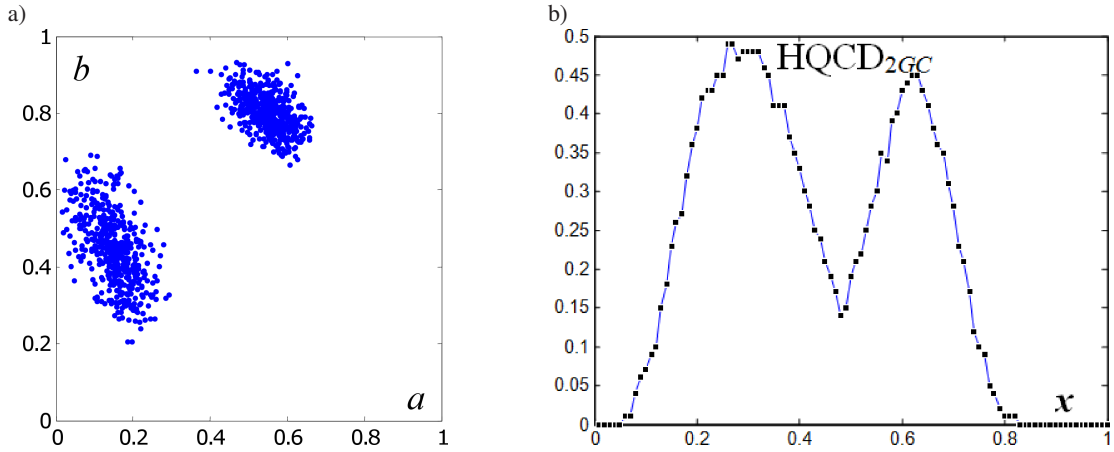

Fig. 14. a) sample set of query boundaries generated using $f_{2GC}(a, b)$, b) HQCD$_{2GC}$ histogram estimating the include function $P_I(y, f_{2GC})$ based on the samples (shown in Fig. 14a)

6.4. 2GC query boundaries distribution – 2 Gaussian Clusters distribution. The probability density function of *2GC* distribution of query boundaries – $f_{2GC}(a, b)$ is a superposition of two Gaussian clusters:

$$f_{2GC}(a, b) = 0.5 \text{ PDF}(\text{N}(m_1, \Sigma_1)) + 0.5 \text{ PDF}(\text{N}(m_2, \Sigma_2)) \tag{17}$$

with mean vectors and covariance matrices:

$$m_1 = \begin{bmatrix} 0.15 \\ 0.45 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 0.0025 & -0.025 \\ -0.0025 & 0.01 \end{bmatrix},$$
$$m_2 = \begin{bmatrix} 0.55 \\ 0.8 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 0.0025 & -0.00125 \\ -0.00125 & 0.0025 \end{bmatrix} \tag{18}$$

is shown in Fig. 13.

The sample set of query boundaries is presented in Fig. 14a. The size of the set is $N_{qb} = 1000$. The set was generated according to the distribution described by $f_{2GC}(a, b)$. HQCD was created using this set. The number of intervals of HQCD equals $N_b = 100$. The HQCD$_{2GC}$ estimating the $P_I(y, f_{2GC})$ is shown in Fig. 14b.

6.5. IU query boundaries distribution – Individual Uniform distribution. (a, b) pairs of query condition endpoints (satisfying *IU* probability distribution) are generated as follows:

- a variable satisfies the unit rectangular uniform distribution (i.e.: a = RAND()),
- b value for a given value of a can be obtained from: $b = (1 - a)$ RAND() + a

were RAND() is a function generating pseudorandom numbers satisfying continuous uniform distribution (rectangular distribution) on the unit interval.

The marginal probability density function of a distribution is given by:

$$f_a(a) = \begin{cases} 1 & \text{for} \quad 0 \le a \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

and it is shown in Fig. 15a.

The conditional probability density function of b distribution for a given value of a (where $0 \le a \le 1$):

$$f_b(b|\, a) = \begin{cases} \dfrac{1}{1-a} & \text{for} \quad a < b \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

and it is shown in Fig. 15b.

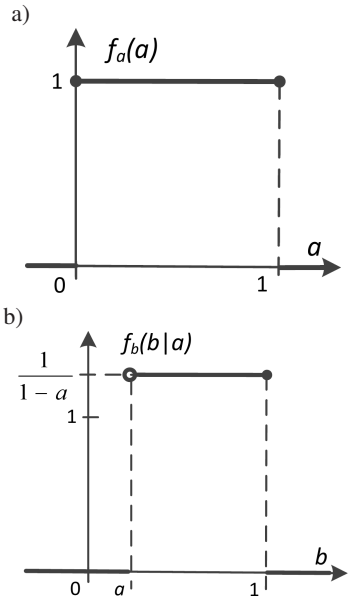

Fig. 15. a) marginal probability density function of a distribution, b) conditional probability density function of b distribution for given value of $a = 1/4$

Using formula for obtaining the conditional probability density:

$$f_b(b|\,a) = \frac{f_{IU}(a,b)}{f_a(a)} \tag{21}$$

we can find the joint density function:

$$\begin{aligned} f_{IU}(a,b) &= f_b(b|\,a) f_a(a) \\ &= \frac{1}{1-a}\left\{\mathbf{1}(b-a) - \mathbf{1}(b-1)\right\}\left\{\mathbf{1}(a) - \mathbf{1}(a-1)\right\}, \end{aligned} \tag{22}$$

where $\mathbf{1}(x)$ is Heavide step function.

The smoothed 2-dimensional histogram – an estimator of f_{IU} based on the samples from the set of query condition boundaries – is presented in Fig. 16a. The sample set is presented in Fig. 16b. The size of the sample set equals $N_{qb} = 1000$.

HQCD histogram based on some sample set (shown in Fig. 16b) is presented in Fig. 17a. The number of buckets of HQCD equals $N_b = 100$. HQCD_{IU} is an estimator $P_I(y, f_{IU})$.

$P_I(y, f_{IU})$ can be obtained from:

$$P_I(y, f_{IU}) = \int_0^y \int_y^1 \frac{1}{1-a} db da = -\ln(1-y)(1-y) \tag{23}$$

and it is shown in Fig. 17b.

We can see a similarity between Fig. 17a and 17b.

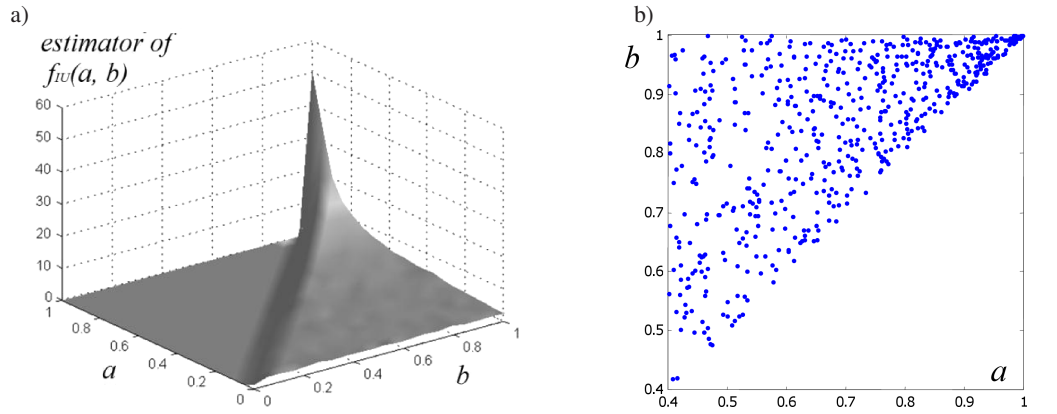

Fig. 16. a) smoothed histogram estimating $f_{IU}(a,b)$ based on a sample set of (a,b) pairs, b) corresponding sample set presented in $[0,1]^2$ space

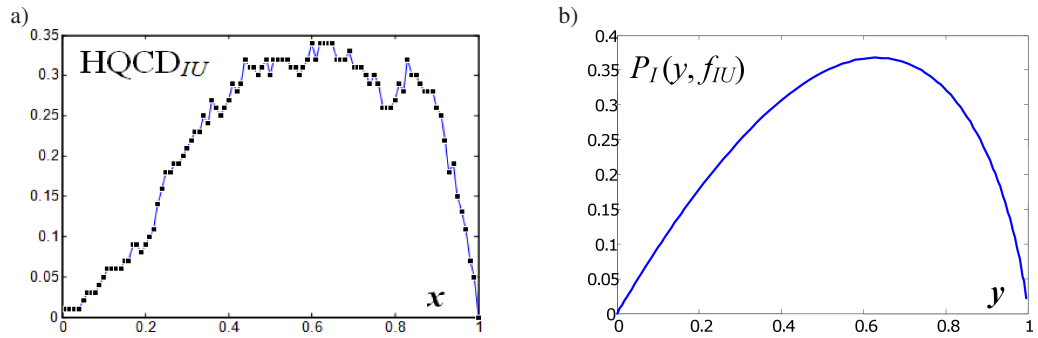

Fig. 17. a) HQCD_{IU} histogram estimating $P_I(y, f_{IU})$ based on samples shown in Fig. 16a, b) include function $P_I(y, f_{IU})$ obtained using the analytical technique

7. Method of validation and experimental results

A relative error metric was used to validate the proposed method of selectivity estimation.

The relative error of selectivity estimation for some Q query is defined as follows:

$$\mathrm{RESE}(Q) = \frac{|sel(Q) - \widehat{sel}(Q)|}{sel(Q)} 100\%, \tag{24}$$

where $sel(Q)$ is an exact selectivity value calculated using F and $\widehat{sel}(Q)$is an approximate selectivity value calculated using QCA-V-optimal histogram (with B buckets) or V-optimal (with B buckets) one or Equi-height one (with $2B$ buckets).

The mean relative error of selectivity estimation is defined as follows:

$$\mathrm{MRESE} = \underset{j=1\ldots N_{qb}}{\mathrm{Avg}} (\mathrm{RESE}(Q_j)) \tag{25}$$

and it is an error metric for a set of queries. The size of the set of conditional endpoints pairs equals N_{qb}. In our experiments we used $N_{qb} = 1000$.

7.1. Experimental results for NI query boundaries distribution.

Using the data sets described in Subsec. 6.1 and 6.2 (*NI* distribution) we obtained the X domain division presented in Fig. 18.

The division presented in Fig. 18 was used to build a QCA-V-optimal histogram. The QCA-V-optimal histogram (bold line) and the standard V-optimal one (dashed line) are shown in Fig. 19.

In Fig. 20 the QCA-V-optimal histogram (bold line) and the HQCD_{NI} one (dashed line) are shown together. This allows to see dependency between regions with high resolution of QCA-V-optimal and regions with high values of HQCD.

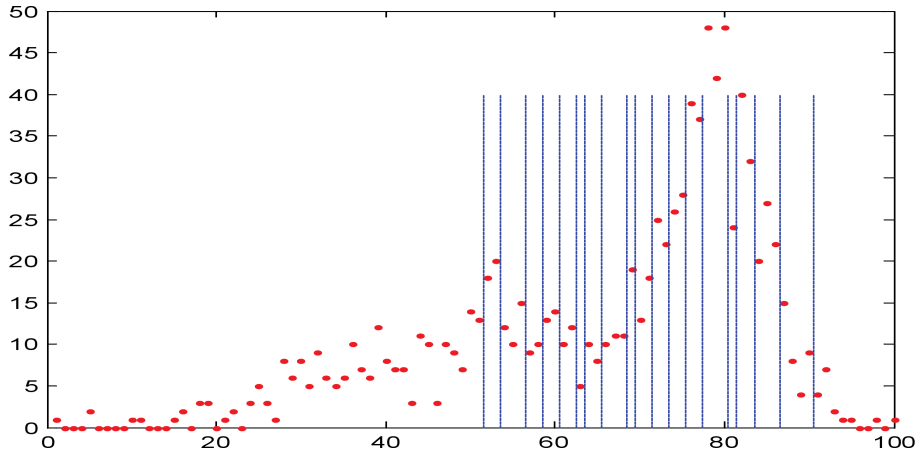

Fig. 18. Error optimal X domain division based on HQCD_{NI} (from Fig. 10b) using SSEW error metric given by Eq. (11). V elements (shown in Fig. 5b) are grouped into buckets of QCA-V-optimal histogram (shown in Fig. 19)

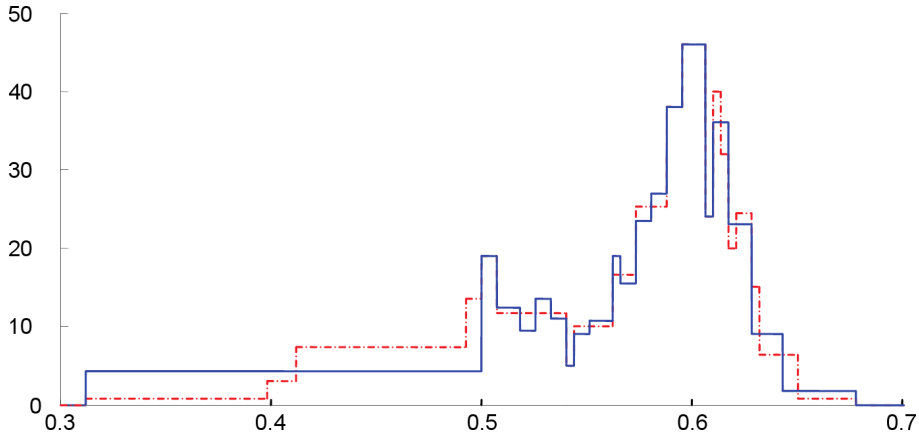

Fig. 19. QCA-V-optimal histogram based on HQCD_{NI} (bold line) shown along the X distribution support; for comparison the classical V-optimal histogram represented by the dashed line is also shown (this V-optimal histogram is individually shown in Fig. 7)

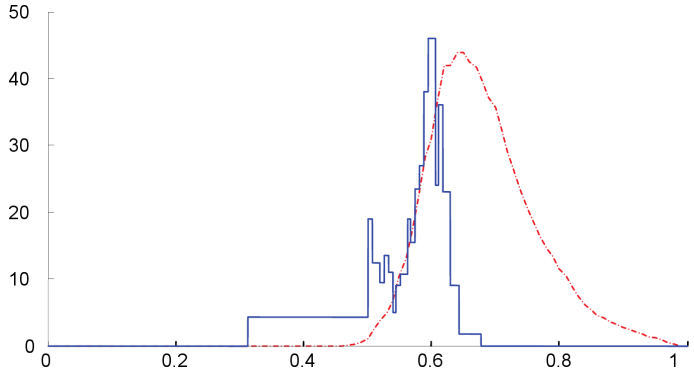

Fig. 20. QCA-V-optimal histogram (bold line) based on HQCD_{NI} shown along the full domain of X, i.e. [0, 1]; the dashed line shows a shape of the relevant HQCD_{NI} histogram (values of the histogram from Fig. 10b are multiplied by 100)

We experimentally obtained mean relative errors for each method of selectivity estimation as follows:

MRESE for QCA-V-optimal histogram $\approx 21.9\%$,

MRESE for V-optimal histogram $\approx 42.0\%$,

MRESE for Equi-height histogram $\approx 37.1\%$.

7.2. Experimental results for 1GC query boundaries distribution. Using the data sets described in Subsec. 6.1 and 6.3 ($1GC$ distribution) we obtained the X domain division presented in Fig. 21.

The division presented in Fig. 21 was used to build a QCA-V-optimal histogram. The QCA-V-optimal histogram (bold line) and the standard V-optimal one (dashed line) are shown in Fig. 22. In Fig. 22 we can see a high improvement of QCA-V-optimal histogram resolution in the middle of X domain range (comparing to the standard V-optimal histogram). Of course, this was achieved at the expense of resolution decrease at the beginning and the end of the X domain range.

In Fig. 23 the QCA-V-optimal histogram (bold line) and the HQCD_{1GC} one (dashed line) are shown together.

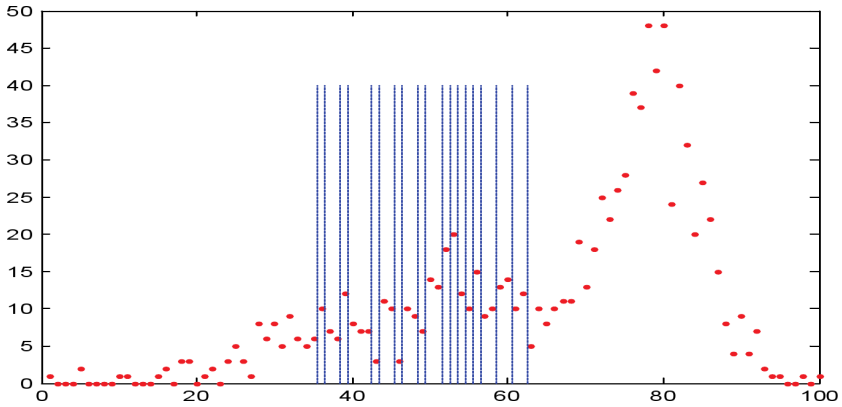

Fig. 21. Error optimal X domain division based on HQCD_{1GC} (from Fig. 12b) using SSEW error metric given by Eq. (11). V elements (shown in Fig. 5b) are grouped into buckets of QCA-V-optimal histogram (shown in Fig. 22)

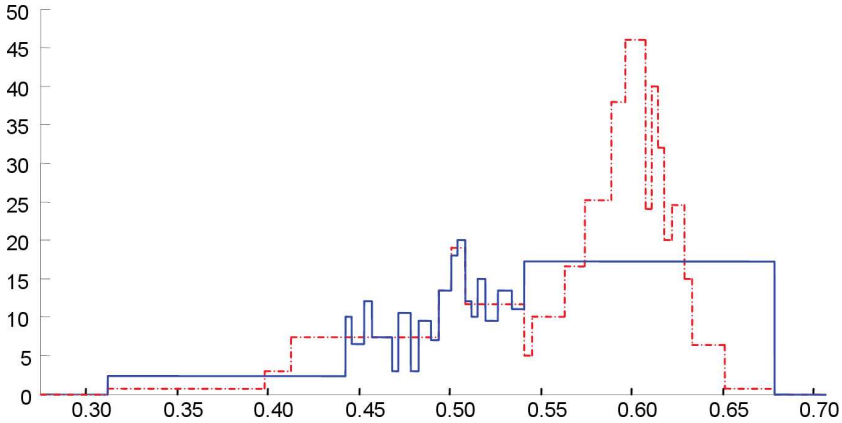

Fig. 22. QCA-V-optimal histogram based on HQCD_{1GC} (bold line) shown along the X distribution support; for comparison the classical V-optimal histogram represented by the dashed line is also shown (this V-optimal histogram is individually shown in Fig. 7)

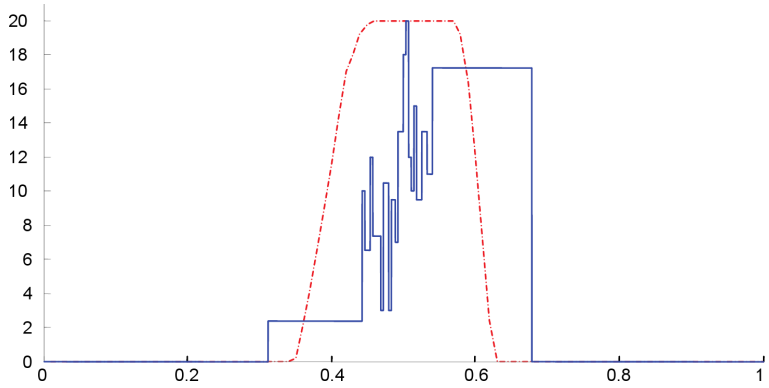

Fig. 23. QCA-V-optimal histogram (bold line) based on HQCD_{1GC} shown along the full domain of X, i.e. [0, 1]; the dashed line shows a shape of the relevant HQCD_{1GC} histogram (values of the histogram from Fig. 12b are multiplied by 20)

We experimentally obtained mean relative errors for each method of selectivity estimation as follows:

MRESE for QCA-V-optimal histogram $\approx$ 15.1%,

MRESE for V-optimal histogram $\approx$ 35.2%,

MRESE for Equi-height histogram $\approx$ 27.1%.

7.3. Experimental results for 2GC query boundaries distribution. Using the data the sets described in Subsec. 6.1 and 6.4 (2*GC* distribution) we obtained the X domain division presented in Fig. 24.

The division presented in Fig. 24 was used to build a QCA-V-optimal histogram. The QCA-V-optimal histogram (bold line) and the standard V-optimal one (dashed line) are shown in Fig. 25.

In Fig. 26 we can see the improvement of QCA-V-optimal histogram resolution at the beginning of X domain range (comparing to the standard V-optimal histogram). Of course this was achieved at the expense of resolution decrease in the middle of X domain range.

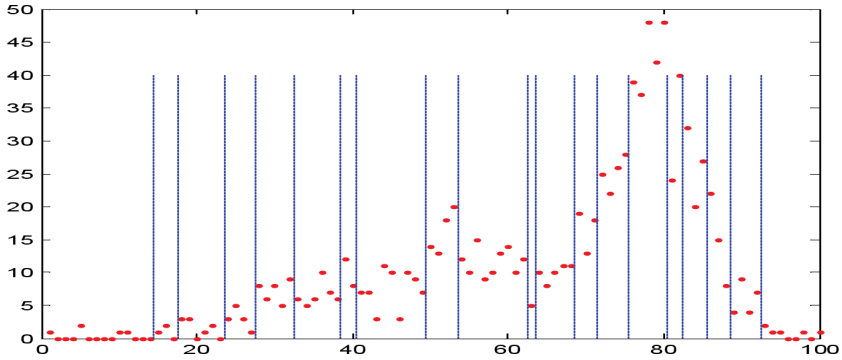

Fig. 24. Error optimal X domain division based on HQCD_{2GC} (from Fig. 14b) using SSEW error metric given by Eq. (11). V elements (shown in Fig. 5b) are grouped into buckets of QCA-V-optimal histogram (shown in Fig. 25)

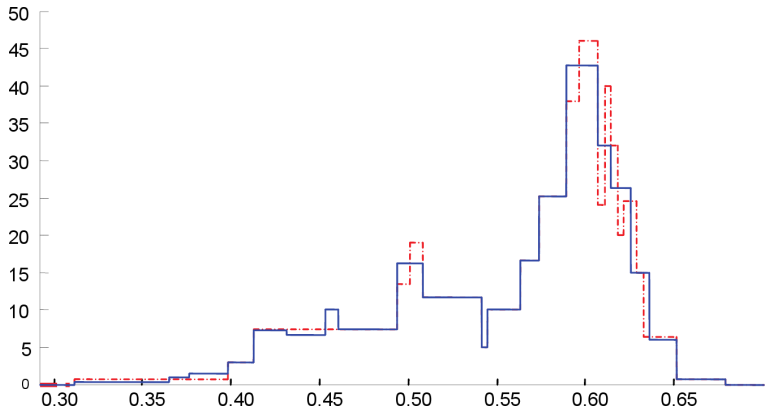

Fig. 25. QCA-V-optimal histogram based on HQCD_{GC} (bold line) shown along the X distribution support; for comparison the classical V-optimal histogram represented by the dashed line is also shown (this V-optimal histogram is individually shown in Fig. 7)

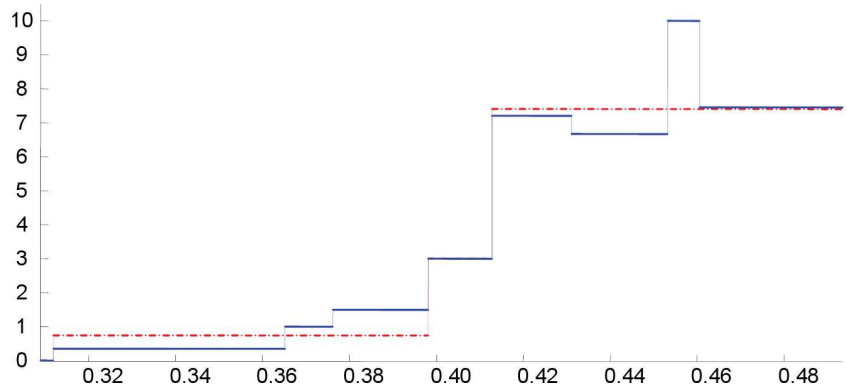

Fig. 26. Zoomed part of QCA-V-optimal histogram (based on HQCD_{2GC} (bold line)) and zoomed part of V-optimal histogram. (The higher resolution of QCA-V-optimal histogram for small values of X (i.e. for high values of HQCD_{2GC}, too))

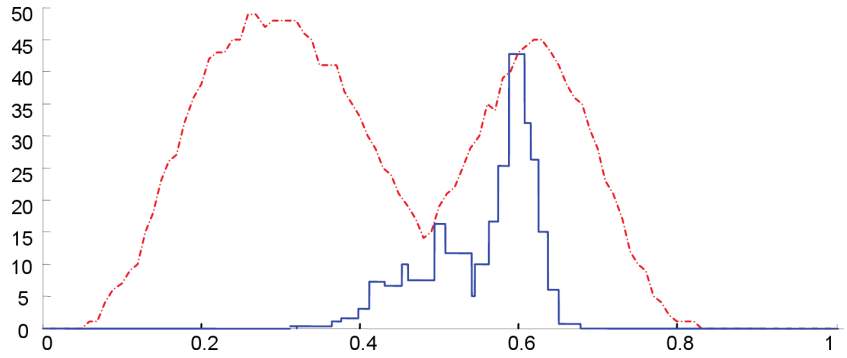

Fig. 27. QCA-V-optimal histogram (bold line) based on HQCD_{2GC} shown along the full domain of X, i.e. [0, 1]; the dashed line shows a shape of the relevant HQCD_{2GC} histogram (values of the histogram from Fig. 14b are multiplied by 100)

In Fig. 27 the QCA-V-optimal histogram (bold line) and the HQCD_{2GC} one (dashed line) are shown together.

We experimentally obtained mean relative errors for each method of selectivity estimation as follows:

MRESE for QCA-V-optimal histogram $\approx$ 26.4%,

MRESE for V-optimal histogram $\approx$ 37.2%,

MRESE for Equi-height histogram $\approx$ 34.0%.

7.4. Experimental results for IU query boundaries distribution. Using the data sets described in Subsec. 6.1 and 6.5 (*IU* distribution) we obtained the X domain division presented in Fig. 28.

The division presented in Fig. 28 was used to build a QCA-V-optimal histogram. The QCA-V-optimal histogram (bold line) and the standard V-optimal one (dashed line) are shown in Fig. 29. In Fig. 29 we can see some increasing of QCA-V-optimal histogram resolution at the beginning of X domain range ($X \in$[0.3, 0.49]), comparing to the standard V-optimal histogram.

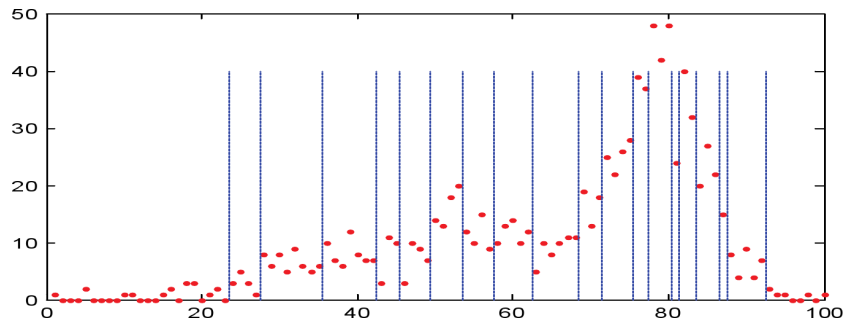

Fig. 28. Error optimal X domain division based on HQCD_{IU} (from Fig. 17a) using SSEW error metric given by Eq. (11). V elements (shown in Fig. 5b) are grouped into buckets of QCA-V-optimal histogram (shown in Fig. 29)

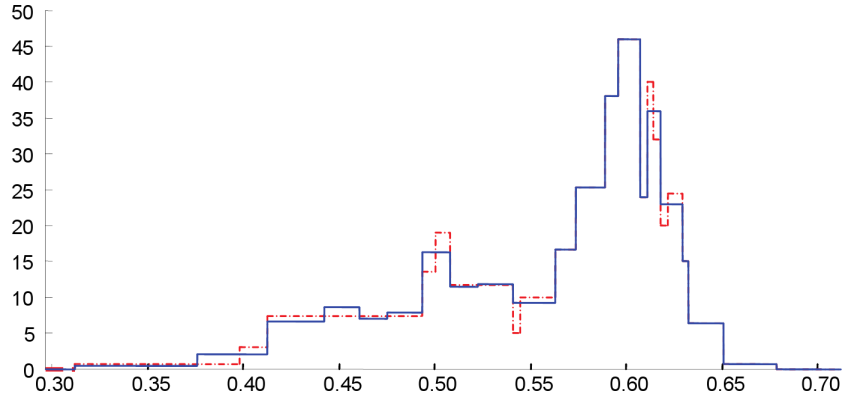

Fig. 29. QCA-V-optimal histogram based on HQCD_{IU} (bold line) shown along the X distribution support; for comparison the classical V-optimal histogram represented by the dashed line is also shown (this V-optimal histogram is individually shown in Fig. 7)

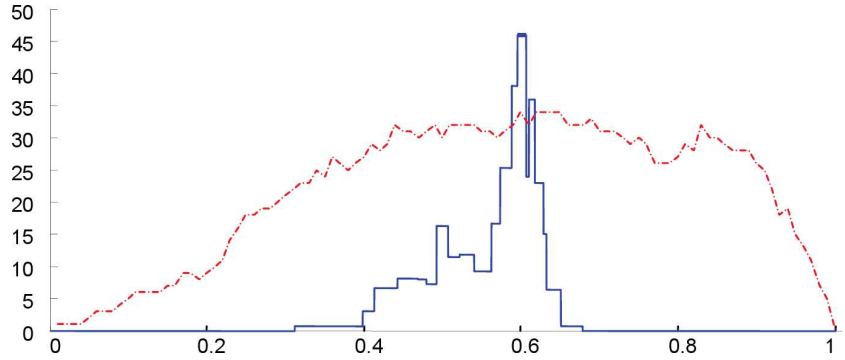

Fig. 30. QCA-V-optimal histogram (bold line) based on HQCD_{IU} shown along the full domain of X, i.e. [0, 1]; the dashed line shows a shape of the relevant HQCD_{IU} histogram (values of the histogram from Fig. 17a are multiplied by 100)

In Fig. 30 the QCA-V-optimal histogram (bold line) and the HQCD_{IU} one (dashed line) are shown together.

We experimentally obtained mean relative errors for each method of selectivity estimation as follows:

MRESE for QCA-V-optimal histogram $\approx$ 16.1%,

MRESE for V-optimal histogram $\approx$ 23.2%,

MRESE for Equi-height histogram $\approx$ 21.9%.

7.5. Experimental results – summary. To obtain statistically significant results we preformed experiments for many instances of set of query interval endpoint pairs. The average of mean relative error of selectivity estimation (AvgMRESE) and the standard deviation of the mean relative error of selectivity estimation (STDMRESE) were experimentally obtained. The number of set instances (generated in accordance with *NI* or 1*GC* or 2*GC* or *UI* distribution) was assumed as 10. Each time a new sample set of X values was generated too (in accordance with the distribution given by Eq. (12)) so a new F vector was also created each time.

The results are presented in Table 1 and Fig. 31. Applying QCA-V-optimal histograms gives the best results for all experiments. For all used query conditional boundaries distributions (i.e. *NI,* 1*GC,* 2*GC, UI*) we obtained the smallest AvgMRESE when QCA-V-optimal histogram was used.

Table 1

Averages and standard deviations of mean relative selectivity estimation for query boundaries distributions: *NI*, 1*GC*, 2*GC, UI* with usage of histograms: QCA-V-optimal, V-optimal, Equi-height

Histogram	MRESE – Mean Relative Error of Selectivity Estimation [%]							
	NI		1*GC*		2*GC*		*UI*	
	Avg	STD	Avg	STD	Avg	STD	Avg	STD
QCA-V-optimal	22.8	1.82	15.2	1.8	27.1	1.52	16.4	1.52
V-optimal	42.12	1.93	35.32	2	37.1	1.8	23.1	1.8
Equi-height	36.5	1.7	28	1.7	32.5	1.75	22.2	1.75

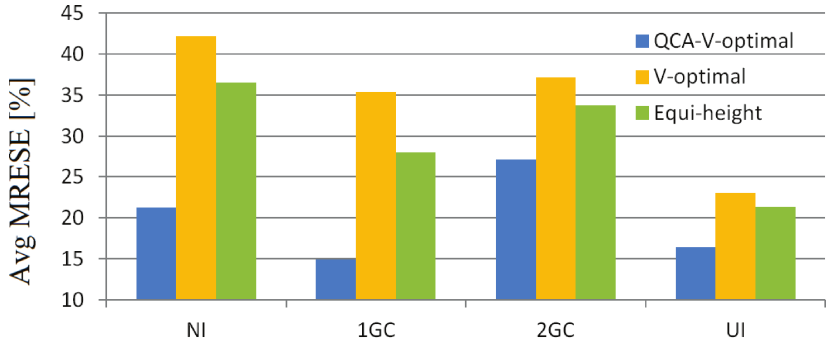

Fig. 31. Averages of mean relative selectivity estimation (AvgMRESE) for query boundaries distributions: *NI*, 1*GC*, 2*GC*, *UI* with usage of histograms: QCA-V-optimal, V-optimal, Equi-height

AvgMRESE (values in columns titled "Avg") in Table 1 differ for QCA-V-optimal, V-optimal, Equi-high. Those differences are statistically significant. For example, using Student's t-distribution we checked that the null hypothesis – "AvgMRESE values are equal for QCA-V-optimal histogram and V-optimal one" should be rejected with some assumed confidence level (value 0.99 was assumed here). Thus, we accepted the alternative hypothesis – "AvgMRESE for V-optimal histogram is greater than AvgMRESE for QCA-V-optimal histogram". In such way, for all exemplary distributions (*NI*, 1*GC*, 2*GC*, *UI*) we accepted the hypothesis that values of AvgMRESE for QCA V-optimal histogram are statistically significantly less then relevant values of AvgMRESE for V-optimal one or Equi-height one (assuming the confidence level equals 0.99).

8. Conclusions and future work

The experimental results (presented in Table 1 and Fig. 31) show that applying a QCA-V-optimal histograms gives the smallest average mean relative error of selectivity estimation (AvgMRESE) for data sets described in Sec. 6. The method based on a QCA-V-optimal-histogram is the best for all used sample sets. This results are rather interesting taking into account that a 1-dimensional HQCD only approximately describes a 2-dimensional distribution of range query boundaries.

Benefits of the proposed method depend on relative location of X domain intervals, i.e.:

- intervals where P_I values are large,
- intervals where F values change significantly.

The overlapping of the mentioned-above intervals is rather required. It is obvious that for some relative locations of intervals this method may be very beneficial.

There is a problem if a description of condition boundaries distribution given by include function is really constant in time. We cannot always assume that P_I does not depend on time (mostly, there are situations in a real system that query condition boundaries distribution may vary in time). The general naïve solution (but which eliminates this problem) is a frequent updating the statistics (i.e. updating HQCD). But we may propose some extension of the method which may also help during the time between sequential updates of HQCD. We propose some simple "method of prediction" of P_I described below.

Let us introduce an uncertain level parameter denoted by p, where $p \in [0, 1)$. Let assume that $f(a, b)$ is a query condition boundaries distribution valid during the time of gathering statistics (during updating a relevant HQCD). Let $f'(a, b)$ denote an unknown query condition boundaries distribution in future. Instead of assuming that query condition boundaries distribution will not change in future (i.e. $f'(a, b) \equiv f(a, b)$) we may predict as follows:

$$f'(a,b) = (1-p)f(a,b) + pf_{2D-uniform}(a,b), \tag{26}$$

where p is a small value (e.g. $p = 0.1$). $f'(a, b)$ becomes a superposition of two distributions. Usage of $f_{2Duniform}$ (see Fig. 2a) allows to include some pair of (a, b) even if they might be eliminated by $f(a, b)$ (i.e. such (a, b) pairs where $f(a, b) \equiv 0$). Applying such approach is easy because of the mentioned linear property of P_I. Using Eq. (26) and (8) we can obtain $P_I(y, f')$ as follows:

$$P_I(y,f') = (1-p)P_I(y,f) + pP_I(y,f_{2D-uniform}). \tag{27}$$

Using Eq. (27) and (10) we can obtain:

$$P_I(y,f') = (1-p)P_I(y,f) + p2y(1-y). \tag{28}$$

The Eq. (28) allows to find a predicted unknown HQCD' (the estimator of $P_I(y, f')$) using known HQCD (the estimator of $P(y, f)$). Let h_j denote a value of HQCD histogram in the j-th bucket defined by interval (b_j, e_j). Analogously, let h'_j denote a value of HQCD' histogram in the j-th bucket defined by interval (b_j, e_j). Thus using Eq. (28) we can find h'_j value as follows:

$$h'_j = (1-p)h_j + 2p \underset{b_j \le y \le e_j}{\text{Avg}} (y(1-y)), \tag{29}$$

where

$$\underset{b_j \le y \le e_j}{\text{Avg}} (y(1-y)) = \frac{1}{e_j - b_j} \int_{b_j}^{e_j} y(1-y)dy = \frac{2b_j^3 - 3b_j^2 - 2e_j^3 + 3e_j^2}{6(e_j - b_j)}. \tag{30}$$

Future work may concentrate on further experimental verification of this approach and development of a method of

choosing the optimal p values. We expect that p value should depend on a level of P_I change obtained during a sequence of time moments between updating of statistics (where the level of P_I change may be estimated as a distance between subsequently obtained HQCD histograms).

Some future work may also concentrate on applying the proposed method in selected DBMSs. Many DBMSs support an interface for extending the functionality of selectivity estimation. For example Oracle DBMS allows to extend CBQO functionality by introducing Oracle ODCI Stat module [15]. This module lets Java or PL/SQL programmers implement a customized representation of an attribute values distribution and a customized method of selectivity estimation (e.g. [13, 16]). It may be used for either creating HQCD histogram (i.e. collecting information about range query boundaries distribution) or creating QCA-V-optimal histogram or enabling (to CBQO) the implementation of selectivity calculating procedure based on QCA-V-optimal histogram.

REFERENCES

[1] Y.E. Ioannidis, "The history of histograms (abridged)", *Proc. VLDB Conf.* 1, CD-ROM (2003).

[2] V. Possala and Y.E. Ioannidis, "Selectivity estimation without the attribute value independence assumption", *Proc. 23rd Int. Conf. on Very Large Databases* 1, 486–495 (1997).

[3] D. W. Scott and S. R. Sain, "Multi-dimensional density estimator", *Handbook of Statistics* 24, 229-263 (2004).

[4] D. Gunopulos, G. Kollios, V.J. Tsortas, and C. Domeniconi, "Selectivity estimator for multidimensional range queries over real attributes", *VLDB J.* 14 (2), 137–154 (2005).

[5] F. Korn, T. Johnson, and H. V. Jagadish, "Range selectivity estimation for continuous attributes", *Proc. Int. Conf. on Scientific and Statistical Database Management* 1, 244–253 (1999).

[6] L. Getoor, B. Taskar, and D. Köller, "Selectivity estimation using probabilistic models", *Proc. ACM SIGMOD Int. Conf. on Management of Data* 30 (2), 461–472 (2001).

[7] L. Lee, K. Deok-Hwan, and Ch. Chin-Wan, "Multi-dimensional selectivity estimation using compressed histogram estimation information", *Proc. 1999 ACM SIGMOD Int. Conf. on Management of Data* 28 (2), 205–214 (1999).

[8] F. Yan, W.C. Hou, Z. Jiang, C. Luo, and Q. Zhu, "Selectivity estimation of range queries based on data density approximation via cosine series", *Data & Knowledge Engineering* 63 (3), 855–878 (2007).

[9] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim, "Approximate query processing using wavelets", *VLDB J.* 10 (2–3), 199–223 (2001).

[10] N. Bruno, S. Chaudhuri, and L. Gravano, "ST Holes: a multidimensional workload-aware histogram", *Proc. 2001 ACM SIGMOD Int. Conf. on Management of Data* 30 (2), 211–222 (2001).

[11] D. Fuchsa, Z. Zhen Heb, and B.S. Lee, "Compressed histograms with arbitrary bucket layouts for selectivity estimation", *Information Sciences* 177 (3), 680–702 (2007).

[12] A. Khachatryan, E. Müller, Ch. Stier, and K. Böhm, "Sensitivity of self-tuning histograms: query order affecting accuracy and robustness", *Proc. 24th Int. Conf. on Scientific and Statistical Database Management* 1, 334–342 (2012).

[13] D.R. Augustyn, "Query-condition-aware histograms in selectivity estimation method", *Proc. Man-machine interactions 2. Advances in Intelligent and Soft Computing* 103, 437–446 (2011).

[14] H.V. Jagadish, V. Poosala, N. Koudas, K. Sevcik, S. Muthukrishnan, and T. Suel, "Optimal histograms with quality guarantees", *Proc. 24rd Int. Conf. on VLDB* 1, 275–286 (1998).

[15] "Oracle 10g documentation, Using extensible optimizer page", http://download.oracle.com/docs/cd/B14117_01/appdev.101/b10800/dciextopt.htm (2005).

[16] D.R. Augustyn, "Applying advanced methods of query selectivity estimation in oracle DBMS", *Proc. Man-Machine Interactions, Advances in Intelligent and Soft Computing* 59, 585–593 (2009).

[17] J. Klamka, K. Grochla, and T. Czachórski, "Modelling TCP connection in WIMAX network using fluid flow approximation", *Proc. 2011 IEEE/IPSJ Int. Symp. on Applications and the Internet. Future Internet Engineering* 1, 502–507 (2011).

[18] D.R. Augustyn, "Application of prediction of atribute value distribution in order to improve accuracy of estimation of question selectivity", *Studia Informatica* 34 2A(111), 23–42 (2013), (in Polish).

[19] D.R. Augustyn, "Using the model of continuous dynamical system with viscous resistance forces for improving distribution prediction based on evolution of quantiles", *Proc. 10th Int. Conf. Beyond Databases, Architectures, and Structures. Communications in Computer and Information* 424, 1–9 (2014).

[20] J. Klamka and J. Tańcula, "Examination of robust stability of computer networks", *Proc. 6-th Working Int.Conf. HET-NETs 2010* 1, 75–88 (2010).

[21] M. Luzar, Ł. Sobolewski, W. Miczulski, and J. Korbicz, "Prediction of corrections for the Polish time scale UTC(PL) using artificial neural networks", *Bull. Pol. Ac.: Tech.* 61 (3), 589–594 (2013).