An Information Theoretic Histogram for Single Dimensional Selectivity Estimation

Chris Giannella and Bassem Sayrafi*

Department of Computer Science, Indiana University Bloomington, IN 47405 USA

E-mail: {cgiannel,bsayrafi}@cs.indiana.edu

Abstract: We study the problem of one dimensional selectivity estimation in relational databases. We introduce a new type of histogram based on information theory. We compare our histogram against a large number of other techniques and on a wide array of datasets. We observe the entropy histograms to fare well on real data. While they do not outperform all methods on all datasets, neither do any other methods. The entropy histograms outperformed all other methods on 4 out of 9 real datasets and tied for first on another two. This conclusion demonstrates that the entropy histograms are an excellent choice of summary structure for selectivity estimation with respect to the state-of-the-art. We also observe that all methods demonstrate a wide variety of behavior across real and synthetic datasets. Along these lines we observe results not consistent with many conclusions drawn in the literature concerning method

*Both authors were supported by NSF Grant IIS-0082407.

accuracy ranking. We believe that the literature has not adequately characterized the performance of previous techniques.

1 Introduction

The problem of creating compact summaries of data has received a significant amount of attention in the relational database field over the last 15 years. The primary application of this work lies in query optimization. Here summaries are used to estimate the result size of relational algebra operations (*e.g.* select, join). These estimates are used in the search for as efficient an SQL query execution plan as feasible ([4]). In particular the problem of estimating the result size of a select operator has received considerable attention (called the selectivity estimation problem). In addition to query optimization, summaries are also used to provide approximate answers to queries [12].

While select operators can involve multiple columns, the selectivity estimation problem for one attribute has received quite a lot of attention in the database literature. In particular, the selectivity estimation problem for queries of the following form has been studied extensively: $\sigma_{a \leq X \leq b}$ where X is a single numerical attribute. This is called the one dimensional range selectivity estimation problem (henceforth “range” is dropped). One reason that this problem has received attention (and continues to do so) is that many commercial DBMS optimizers still make heavy use of single dimensional selectivity estimation.

1.1 Contributions

We introduce a new type of histogram based on information theory. The key idea is that the quality of a bucket is the degree to which its frequency distribution differs from uniform. We use the Kullbach-Leibler distance to quantify the degree of difference from uniformity. Then bucket boundaries are chosen to maximize a sum of the quality of all buckets with each weighted by its number of distinct values. The result is two novel histograms (based on frequencies and areas [13])

We compare the entropy histogram against a large number of other techniques. To our knowledge, our study represents the largest comparison of one dimensional selectivity estimation techniques in recent times (since 1996 [13]).[1] We compare across a larger number of real datasets (9) than any other study.

From an experimental stand-point, our contribution is two-fold. First we evaluate the performance (accuracy) of the entropy histograms against all other methods. We demonstrate that these histograms represent an excellent choice of summary structure for selectivity estimation with respect to the state-of-the-art. Note that we did not observe the entropy histograms to outperform all other methods on all real dataset. However, neither do we observe any other method to do so. Nevertheless, the entropy histograms have superior performance on more real datasets than any other method.

The observation that no method is superior on all datasets leads to our second experimental contribution. We observe that all methods demonstrate a wide variety of behavior across real and synthetic datasets. For example, linear wavelets has accuracy tied for first on one real dataset and last on another. Also, MDA and OVA outperforms all other methods on a synthetic dataset, but rank nearly last on a real dataset. Moreover, we also observe results not consistent with many conclusions drawn in the literature. As such we do not believe that the performance of previous methods have be adequately characterized in the literature. Our paper represents a step toward addressing this issue.

1.2 Outline

In Section 2 we lay out our notation and background definitions followed by a formal definition of the problem addressed. In Section 3 we discuss related work. In Section 4 we introduce the optimal entropy

[1] largest in terms of the number of techniques compared

histograms. In Section 5 we describe the experimental setup and preliminary results. In Section 6 we present the results for real and synthetic data. In Section 7 we discuss our results and draw conclusions regarding the relative performance of the entropy histograms and the extent to which past literature has or has not adequately characterized the relative performance of all methods. Finally, in Section 8 we conclude the paper.

2 Notation and Problem Statement

We use notation nearly identical to [13]. Let R be a relational database table and X be an attribute of R with domain $\mathcal{D} \subseteq \mathbb{R}$.[2] Let $\mathcal{V}$ be the active domain of X in R (*i.e.* $\{v \in \mathcal{D} : \exists$ tuple t in $R, t[X] = v\}$). We denote the elements of $\mathcal{V}$ as $v_1, \ldots, v_n$ where $v_i < v_j$ if $i < j$. The frequency f_{v_i} of v_i ($1 \leq i \leq n$) is defined as the number of tuples t in R where $t[X] = v_i$ (note $f_{v_i} > 0$). The cumulative frequency of v_i is $c_{v_i} = \sum_{j=1}^{i} f_{v_j}$. The spread of v_i ($1 \leq i < n$) is defined as $s_i = v_{i+1} - v_i$. The area of v_i is defined as $a_i = f_{v_i} * s_i$. The data distribution of X is $\mathcal{T} = \{(v_i, f_{v_i}) : 1 \leq i \leq n\}$ and the cumulative data distribution of X is $\mathcal{T}^C = \{(v_i, c_{v_i}) : 1 \leq i \leq n\}$. By extending out over the entire domain $\mathcal{D}$, we define the extended cumulative data distribution as $\mathcal{T}^{C+} = \{(v, c_v) : v \in \mathcal{D}\}$. Here, c_v equals zero if $v < v_1$, otherwise c_v equals c_{v_i} where v_i is the largest value in $\mathcal{V}$ such that $v_i \leq v$.

Selection queries are denoted $\sigma_{a \leq X \leq b}$ where $-\infty, \infty$ are permitted for a and b, respectively. The cardinality of a range selection query is defined as the number of tuples t in R such that $a \leq t[X] \leq b$ and is denoted $|\sigma_{a \leq X \leq b}|$ ($a = -\infty$ and $b = \infty$ allow for one-sided queries). A *selectivity class* a sub-interval $[\alpha, \beta] \subseteq [0, 1]$ and $\sigma_{a \leq X \leq b}$ is in the class if $\frac{|\sigma_{a \leq X \leq b}|}{|R|} \in [\alpha, \beta]$.

Problem statement: The one dimensional selectivity estimation problem is defined as follows. Con-

[2] The domain of X is the set of all values which could possibly appear in the X column of R.

struct a compact summary data structure $\mathcal{S}$ from $\mathcal{T}$ (or $\mathcal{T}^C$ or T^{C+}) which can be used to accurately estimate the cardinality of any given range selection query. The heart of the problem lies in the trade-off between compact and accurate – a poorly compact summary structure can achieve excellent accuracy. Indeed, storing $\mathcal{T}$ as a list would enjoy perfect accuracy but the worst possible degree of compaction. Since the summary structure will be stored in the DBMS catalog, it must be quite compact. Hence, the trade-off is made particularly acute.

3 Related Work

The selectivity estimation problem has recently received quite a lot of attention in the database community. The research can be divided into four categories: one dimensional selectivity estimation (non-continuous, static data), multi-dimensional selectivity estimation, selectivity estimation over continuous data, and other facets (*e.g.* summary structure maintenance). We will not discuss the last three categories. For this, we refer the interested reader to a recent survey [8]. We focus on the first category since it is where our proposed method fits. We further categorize the past literature into three groups: piecewise constant histograms, piecewise linear histograms, and others (*i.e.* wavelets and sampling).

Histograms A histogram $\mathcal{H}$ is an m-tuple $\langle B_1, \ldots, B_m \rangle$ where each B_i (called a bucket) is a compact, summary of some $P_i \subseteq \mathcal{T}$ having consecutive data values. $P_1, \ldots, P_m$ form a partition of $\mathcal{T}$ into non-empty, pairwise disjoint sets. $\mathcal{H}$ can be thought of as representing a piecewise approximation of $\mathcal{T}$.

Let $V(P_i) = \{v : \exists(v, f_v) \in P_i\}$ denote the values in P_i and $F(P_i)$ denote the collection of frequencies that appear in P_i (including repeats). To approximate P_i, histograms keep track of the starting and ending values $v(i)_{lo} = min\{v : \exists(v, f_v) \in P_i\}$, $v(i)_{hi} = max\{v : \exists(v, f_v) \in P_i\}$ and some representation of an approximation of the values and frequencies in P_i. All of the histograms we implement and include in our study approximate values using the uniform spread assumption [13]: they assume that the spreads are the same. Hence $v \in V(P_i)$ (assumed to be the j^{th} value $1 \leq j \leq |V(P_i)|$) is approximated as $v' = v(i)_{lo} + (j-1)\frac{v(i)_{hi} - v(i)_{lo}}{|V(P_i)|-1}$. Let f'_v denote the approximation to f_v used by any of the histogram methods. Several different approximations are used and will be described below.

Given a query $q = \sigma_{a \leq X \leq b}$, let S_q denote its selectivity $|\sigma_{a \leq X \leq b}|$. All histograms we implemented and included in our study approximate S_q by $S'_q = \sum_{i=1}^{m} \sum \{f'_v : v \in V(P_i), a \leq v' \leq b\}$.

To construct $\mathcal{H}$ two questions must be addressed: how is f'_v computed, and how to pick the endpoints for each bucket?

Piecewise constant histograms: Several proposals in the literature approximate the frequencies in P_i by their average (uniform frequency assumption [13]). Each frequency in P_i is approximated as the average frequency $\overline{f_i} = \frac{\sum F(P_i)}{|F(P_i)|}$ (*i.e.* $f'_v = \overline{f_i}$).

All the piecewise constant histograms we consider store the following information in each B_i: $v(i)_{lo}$, $\overline{f_i}$, and $|V(P_i)|$ (the right end-point can be inferred from bucket B_{i+1}). The last bucket is required to store both end-points, hence a total of $3m + 1$ numbers are stored.

Two methods for choosing the bucket boundaries (given a fixed number of buckets m) were described and experimentally evaluated by Poosala *et al.* [13]. The first places bucket boundaries between the $m - 1$ largest consecutive frequency differences $f_{v_{i+1}} - f_{v_i}$. The result is called the *Max-diff histogram*. The second places bucket boundaries so as to minimize the following sum: $\sum_{i=1}^{m} |V(P_i)| Var(P_i)$ where $Var(P_i)$ denotes the variance of the frequencies in P_i.[3] The result is called the *optimal variance histogram*. The first efficient algorithm for constructing optimal variance histograms was found by Jagadish *et al.* [9] (worst case time $O(m * |\mathcal{V}|^2)$). Poosala *et al.* also consider the the *Max-diff area histogram*

[3] Equivalent to minimizing the sum squared error between each frequency and its estimate $\overline{f_i}$.

and optimal variance area histogram (areas are used instead of frequency above).

Gilbert *et al.* [5] develop an algorithm which computes a histogram minimizing the squared difference between the query selectivity and the histogram selectivity estimation summed over all possible range queries over the value set. This is a generalization of the optimal variance histogram which minimizes the squared difference over all *point* queries. This algorithm has high computational complexity (although polynomial), so they also develop a more efficient algorithm which does not guarantee finding an optimal histogram. The result is called the *A0 histogram*. We use a modified version of A0. As mentioned earlier, we store $|V(P_i)|$ in each bucket, but this number is not stored in by Gilbert *et al.*. They assume that the values in $\mathcal{T}$ are a contiguous set of integers and that frequencies may be zero (hence $|V(P_i)| = v(i)_{hi} - v(i)_{lo}$). We do not assume the values are contiguous. We consider only range queries over values appearing in $\mathcal{T}$, when computing the simplified difference squared across all range queries.

Buccafurri *et al.* [2] propose a different method for representing the values and frequencies in a bucket. We did not implement and include this technique in our study because it is orthogonal to the issue to choosing bucket boundaries (their technique could easily used on top of ours and others). Guha *et al.* [6] developed an algorithm for finding a histogram which minimizes the squared difference between the query selectivity and the histogram selectivity estimation summed over a pre-defined collection (workload) of hierarchical range queries. We did not implement and include this technique in our study because our technique (and many others we consider) do not take into account pre-defined workload information, but could be modified to do so. We leave these two additions to future work.

Piecewise linear histograms: König and Weikum [10] develop piecewise linear histograms. A linear regression line is used to approximate the frequencies in P_i The independent variable ranges over values and the dependent over frequencies. Each bucket B_i stores: $v(i)_{lo}$, m_i, b_i, and $|V(P_i)|$ where m_i represents the slope of P_i and b_i the y-intercept. Frequencies are approximated as follows $f'_v = m_i * v' + b_i$ (v' is the value estimate based on the uniform spread assumption). Bucket boundaries are computed to minimize the sum square error over all frequencies and their approximation above using a slight modification of the algorithm in [9]. We call the result the *linear histogram*.

König and Weikum also described a method for using the results from previous queries (feedback) to adjust the histogram. We did not implement and include this technique in our study because it can be added to any of the other histograms without change. We leave such an addition to future work.

Zhang and Lin [16] extend the linear histogram approach. They develop a different way of estimating each bucket with a linear regression line. They observe that keeping a standard linear regression line in each bucket does not guarantee that the estimated count of tuples in a bucket is the same as the actual count or that the estimated sum of all values in the same as the actual sum. They describe how to use both these conditions to calculate the coefficients m_i and b_i (linear spline histogram for summation with count and sum guaranteed). They approximate frequency f_v in the same way as [10]. Bucket boundaries are computed to minimize the quantity $\sum_{i=1}^{m} \sum_{v \in V(P_i)} (f'_v * v - f_v * v)^2$ with f'_v as described above using a slight modification of the technique in [9]. We call the resulting histogram the *spline histogram*. If area is used in place of frequencies, then we call the resulting histogram the *spline area histogram*.

For both the linear histogram and the spline histograms, we stored four numbers for each bucket (five for the last bucket), thus $4m + 1$ numbers in total.

Other methods: Wu *et al.* [15] develop a method based on the so-called “golden rule of sampling”. They generate a sample from $\mathcal{T}^C$ which is used to approximate the selectivity of a given query $q = \sigma_{a \leq X \leq b}$. Each sample point requires two numbers of storage.

Matias *et al.* [11] develop two methods based on Haar and linear wavelets. The wavelet coefficients plus the average of the dataset is computed for the extended cumulative distribution T^{C+}. The smallest coefficients are dropped resulting in m coefficients plus the average of the dataset. Each coefficient requires two numbers of storage, one for the coefficient and one for its position. Hence, $2m+1$ numbers are required in total. We implemented the Haar wavelet method as it is described in [11] and linear wavelet method as described in [14].

4 Information Theoretic Histogram

In this section, we describe a novel type of piecewise constant histogram based on information theory. Our motivation comes from revisiting the idea behind the optimal variance histogram. There a histogram was constructed which minimized expression $\sum_{i=1}^{m}|V(P_i)|Var(P_i)$. This expression can be thought of as measuring the "goodness" of a particular histogram (*i.e.* choice of bucket boundaries). Each term is a measure of the goodness of each bucket. $Var(P_i)$ is the average sum squared error in approximating each individual frequency in P_i by the average frequency over P_i. As such, $Var(P_i)$ can be thought of as measuring the degree to which the frequency distribution differs from uniform. The $|V(P_i)|$ factor can be thought of as weight penalizing more heavily buckets containing larger numbers of distinct values.

We follow a similar intuition but propose a different measure of proximity to uniformity. We consider the relative frequency distribution $RF(P_i) = \{\frac{f}{\sum F(P_i)} : f \in F(P_i)\}$, (repeated elements not dropped). The degree of difference from uniformity of this distribution can be measured using any one of the many techniques for measuring the degree of difference between two probability distributions. One common method is the Kullbach Leibler measure ([3] page 18). When applied to the relative frequency distribution and the uniform distribution over $|V(P_i)|$ elements, the result is $\log_2(|V(P_i)|) - Ent(P_i)$ where $Ent(P_i)$ is the Shannon entropy of the relative distribution for P_i, $-\sum_{p \in RF(P_i)} plog_2(p)$. As above, a $|V(P_i)|$ weighting factor can be used to penalize more heavily buckets containing larger numbers of distinct values. The result is the following expression which measures the goodness of a particular histogram: $\sum_{i=1}^{m}|V(P_i)|(log_2(|V(P_i)|) - Ent(P_i))$.

We call the histogram which minimizes the above expression the *optimal entropy histogram*. A slight modification of the technique from [9] can be used to find the histogram efficiently. We also consider the histogram which results when the above expression with areas used in place of frequencies in $F(P_i)$. We call the resulting histogram the *optimal entropy area histogram*.

5 Experimental Setup and Preliminary Results

5.1 Experimental Setup

For a fixed query q, let S_q denote the number of tuples returned by q. For a given selectivity estimation method, let S'_q denote the estimated number of tuples. We define the *selectivity estimation error* on query q to be $|S_q - S'_q|$. Given a collection of queries $q = 1, \ldots, Q$, we define the *average relative error percentage* (for a given selectivity estimation method on a given dataset) as: $E_{ave} = \frac{100}{Q}\sum_{q=1}^{Q}\frac{|S_q - S'_q|}{S_q}$.

An *experiment* consists of fixing a dataset (synthetic or real), the amount of storage allocated to each selectivity estimation method, the number of queries (Q), and the selectivity class $[\alpha, \beta]$, then randomly generating Q queries from class $[\alpha, \beta]$ and computing E_{ave} (and the relative error variance). In all our experiments Q was fixed at 40000 and $[\alpha, \beta]$ at $[0, 0.2]$.

Datasets: We conduct experiments on both synthetic data and real datasets. Synthetic data was generated according to the model introduced in [13]. The model has five parameters: number of distinct values (n), total number of tuples (N), value range (VR, the difference between the largest and smallest distinct value), spread Zipfian skewness parameter ($spread\_z$), and the frequency skewness parameter ($freq\_z$). We further assume that the spreads follow the cusp_max assumption and that values are randomly assigned to frequencies (full details can be found in [13]).

Several real datasets are used to compare the performance of the selectivity estimation methods: SGIadult ($N = 32561$, $n = 73$, $VR = 73$), shuttle2 (43500, 177, 9896), forestov4 (581012, 551, 1397), forestcov9 (581012, 255, 254), cup199 (95413, 1135, 1500), cup472 (95413, 60, 200), ipums25 (88443, 77, 99), ipums51 (88443, 918, 1009998), and ipums52 (88443, 93, 1009998). All datasets consist of integers; N is the total number of tuples, n the number of distinct values, and VR the value range.

The SGIadult dataset was obtained from the UCI machine learning archive [1] (called the "adult" dataset there). We use the age column of the training dataset. The dataset was extracted from 1994 US census data. The shuttle2 dataset was downloaded from the "Esprit Project 5170 StatLog" archive ("Shuttle" heading): www.liacc.up.pt/ML/. It represents data concerning the operation of the NASA space shuttle. We use attribute two. The remaining datasets were obtained from the UCI KDD archive [7]. The forestcov4 and forestcov9 datasets were found under the "Forest CoverType" heading, covtype.data file – attributes four and nine, respectively. The attributes represent various geographic measurements. The cup199 and cup472 datasets can be found under the "KDD CUP 1998 Data" heading, cup98lrn file – attributes 199 (IC1, median household income) and 472 (TARGET_D, donation amount quantized into 60 groups), respectively. The ipums25, ipums51, and ipums52 datasets were found under the "IPUMS Census Data" heading, ipums.la.99 file – attributes 25 (eldch, age of the eldest child in the household), 51 (incbus represents business income), and 52 (incfarm represents farm income), respectively.

Summary structure sizes: In our experiments we report summary structure size in terms of piecewise constant histogram buckets (labeled "buckets"). An m bucket histogram requires $12m+1$ bytes of space (each bucket requires 3 numbers excepts the last, 4 numbers). Hence, in this space, we can store: an $\lfloor 0.75m \rfloor$ bucket linear or spline histogram; a wavelet structure with $\lfloor 1.5m \rfloor$ coefficients; and a sampling structure with $\lfloor 1.5m + 0.125 \rfloor$ points.

Figure legends: In all figures, the following legend is used:Max-diff histogram and Max-diff area histogram [13], MD and MDA; optimal variance histogram and optimal variance area histogram [13], OV and OVA; optimal entropy histogram and optimal entropy area histogram, OE and OEA; linear histogram [10], LH; spline histogram and spline area histogram [16], SP and SPA; A0 histogram [5], A0; sampling [15], SMP; Haar and linear wavelets [11], HWV and LWV. In some cases, multiple methods are depicted as one curve. In these cases, the difference between the methods was very small. One curve is used to make the figures more readable.

5.2 Preliminary Results

Construction time: For brevity we do not present out timing results in full, but summarize as follows. With the exception of A0, all structures were constructed in less than 45 seconds on datasets with up to 1135 distinct values (cup199). A0, however, required approximately 4000 seconds on forestcov4 (551 distinct values) and was terminated early on cup199 (1135) because it required far too much time. We conclude that the construction times for all methods except A0 are quite reasonable.

Statistical significance: Consider $E1_{ave} > E2_{ave}$, the average error of two methods in an experiment. With only a few exceptions[4], if $E2_{ave}$ is less than 86 percent of $E1_{ave}$, then the average relative error of method 2 is less than method 1 with 90% statistical signifigance. By "eye-balling" the accuracy figures with this in mind, the reader can get a good idea whether the appearance that one method is outperforming another is statistically significant. For bevity, we omit detailed discussion of the technical

[4] cup472, ipums25, ipums51, and ipums52 at 5 buckets

Figure 1. Average relative error for real data sets (15 buckets)

	MD	MDA	OV	OVA	OE	OEA	LH	SP	SPA	A0	SMP	HWV	LWV
forestcov9	205	205	19	19	7	7	21	9	9	14	55	53	7
SGIadult	229	229	16	16	2	3	56	1	1	12	70	54	5
ipums25	898	898	108	108	1	2	64745	5	5	32	68199	98	10
cup199	997	830	140	143	10	21	71	11	11	**	47	58	28
forestcov4	938	1022	215	539	5	6	332	42	42	13	73	54	25
cup472	1094	239	1094	190	21	174	12874	129	129	514	176	173	83
shuttle2	1584	787	1437	110	202	17	729	46	45	492	127	79	166752
ipums51	16	10	16	13	2118	2350	246	3	4	**	2345	627	1547
ipums52	29	2	6	3	13	20	325	1	2	13	157037	64031	99889

justification of the above criteria.

6 Results

6.1 Real Data

Figure 1 depicts a cross-section of the results across all real data sets (the storage structure size fixed at 15 buckets).[5] Figure 2 depicts the accuracy ranking of the methods on each dataset (the leftmost has the highest accuracy, rightmost has the lowest). Methods which could not be separated at the 90% statistical signifigance level using the technique described earlier are grouped together. Figure 3 (top graph) depicts the error for all methods on forestcov4 as the number of buckets varies. The bottom graph depicts the best performing methods.[6]

Entropy histograms performance: From Figure 2, we see that no method is superior in all cases. However, OE ranks ahead of all other methods on 3 of 9 datasets (ipums25, forestcov4, cup472), and

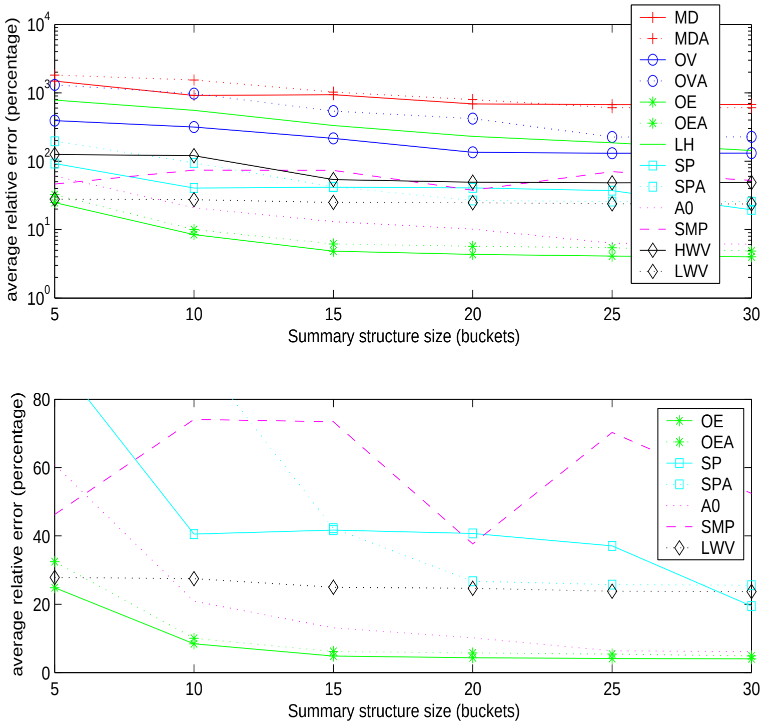
[5] Due to excessive construction times A0 was not used on the cup199 and ipums51 (indicated by **).

[6] For bevity, we do not present graphs for all datasets.

Figure 2. Accuracy rankings: highest on the left (15 buckets).

forestcov9	OE OEA,LWV	SP SPA	A0	OV OVA,LH	SMP HWV	MD MDA					
SGIadult	SP SPA	OE	OEA	LWV	A0	OV OVA	LH HWV	SMP	MD MDA		
ipums25	OE	OEA	SP SPA	LWV	A0	OV OVA HWV	MD MDA	LH SMP			
cup199	OE SP SPA	OEA	LWV	SMP	HWV	LH	OV OVA	MDA	MD		
forestcov4	OE	OEA	A0	LWV	SP,SPA	HWV	SMP	OV	LH	OVA	MD,MDA
cup472	OE	LWV	SP SPA	OVA OEA, SMP HWV	MDA	A0	MD OV	LH			
shuttle2	OEA	SP,SPA	HWV	OVA,SMP	OE	A0	MDA,LH	MD,OV	LWV		
ipums51	SP	SPA	MDA	OVA	MD OV	LH	HWV	LWV	OE OEA SMP		
ipums52	SP	MDA,SPA	OVA	OV	OE,A0	OEA	MD	LH	HWV	LWV	SMP

Figure 3. forestcov4 dataset



is tied for first on forestcov9 and cup199. This type of performance is slightly better than the next best method, SP. It ranks ahead of all others (except its area counterpart) on 3 of 9 (SGIadult, ipums51, ipums52) and is tied for first on one (cup199). If we also consider histograms together with their area counterpart, then the entropy method remains slightly superior to the spline method. OE or OEA ranks ahead on 4 of 9 datasets (tied on 2) and SP or SPA ranks ahead on 3 of 9 (tied on 1). No other method is close in terms of this type of performance.

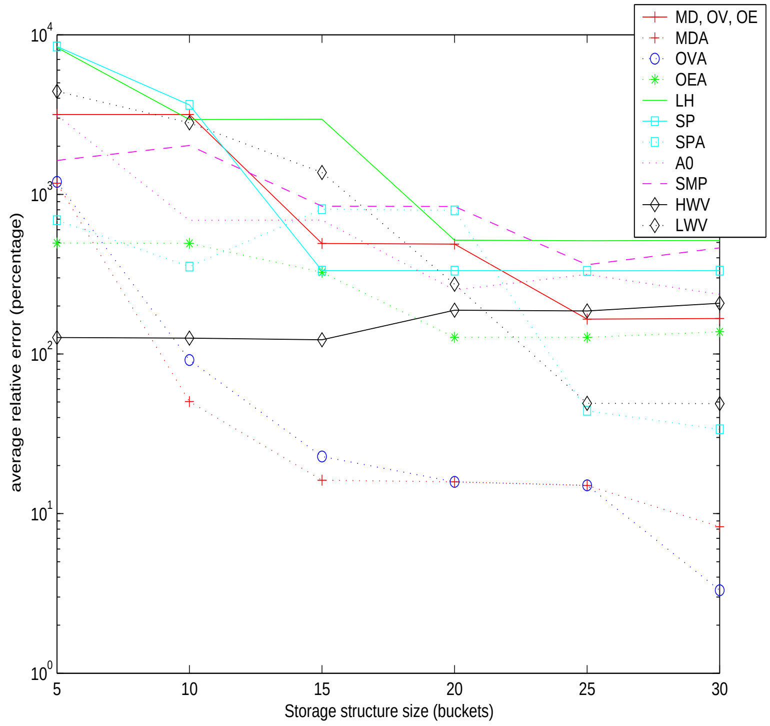
From Figure 1, we can take a closer look at the actual accuracy differences. We see that: OE is 5 times more accurate than SP on ipums25 (although both are very accurate), 8 times more on forestcov4, and 6 times more on cup472. On the other hand, SP is 2 times more accurate on SGIadult (although both are very accurate), 4 times more on shuttle2 (although 3 times less than OEA), 783 times more on ipums51, and 13 times more on ipums52.

From Figure 3, we can get a snapshot of how the number of buckets affects accuracy. As expected the accuracy of all methods (except SMP) increases.[7] OE is at least 4 times more accurate than SP at all numbers of buckets. The closest competitor seems to be A0 (except at 5 buckets). OE is at least twice as accurate for ≤ 20 buckets and approximately equivalent for 25 and 30 buckets.

Performance variations: Interestingly, all methods exhibit substantial variations in accuracy across datasets. OE and OEA are ranked first on several datasets and last on ipums51 (with errors greater than 2000). The variations for SP and SPA are not as extreme: ranked first on several datasets and fifth on forestcov4 with errors of 42. Other methods also exhibit large variations, some examples include the following. MDA (OVA) ranked second (third) on ipums52 with an error of 2 (3) but last (second to last) on forestcov4 with an error of 1022 (539). LWV ranked first on forestcov9 with an error of 7 but last on

[7] Due to the random nature of sampling, it need not always increase.

Figure 4. Synthetic data, $n = 500$, $N = 500000$, $VR = 100000$, $freq\_z = 1$, $spread\_z = 2$



shuttle2 with an error of 166752.

6.2 Synthetic Data

Figure 4 depicts accuracy on synthetic dataset parameters consistent with those used in [13] and reported as “typical performance” (note the y-axis is on a logarithmic scale). Unlike the real datasets, MDA and OVA demonstrate superior performance for all but 5 buckets (particularly at 15 and 20 buckets where they are nearly an order of magnitude more accurate than the next nearest method). OEA, SP, and SPA have mediocre performance – nowhere near the best.

7 Discussion

In this section we discuss the results with respect to (1) the relative performance of the entropy histograms and (2) the extent to which our results match past literature.

7.1 Relative Performance of the Entropy Histograms

On real data, the entropy histograms generally fared very well. OE had the best accuracy on 3 out of 9 datasets and was tied for first on another two. The next closest competitor was SP which had the best accuracy on 3 out of 9 and was tied on another one. Moreover if we consider these histograms together with their area counterparts we conclude that OE or OEA ranks first on 4 out of 9 (tied on another 2) and SP or SPA ranks first on 3 out of 9 (tied on another one).

In conclusion, the entropy histograms generally had excellent performance on real data, but were not superior on all datasets. But neither were any other methods. The entropy histograms were superior on more datasets than any other method. The entropy histogram is an excellent choice of summary structure with respect the the state-of-the-art.

On one real dataset the entropy histograms had an extremely large error, the worst accuracy over all other methods (except sampling). Moreover, on synthetic data, the entropy histograms had mediocre performance, nowhere near the best. This indicates that the relative accuracy of the entropy histograms is data dependent and that in order to assess their performance with respect to the other methods a wide variety of datasets must be used. However, this same conclusion can be drawn about all other methods and will be discussed further next.

7.2 Comparison with Past Literature

Many studies have appeared since 1996 that propose a new selectivity estimation method along with experimental results showing it to outperform many (or all) previous methods. MDA and OVA were proposed in 1996. HWV and LWV were proposed in 1998 and argued to outperform MDA. LH was proposed in 1999 and argued to outperform MDA.[8] SMP was proposed in 2001 and argued to outperform wavelets. A0 was proposed in 2001 and argued to outperform wavelets. Finally, SP and SPA were proposed in 2002 and argued to outperform MDA, OVA, and LH.

Our results show, however, that no definitively best method exists. Moreover, they present a case against *all* of the results described above. On two real datasets and a synthetic dataset we observed MDA to outperform both HWV and LWV. On four real datasets we observed both HWV and LWV to outperform SMP. On one real dataset we observed both HWV and LWV to outperform A0. Finally, on a synthetic dataset we observed both MDA and OVA to outperform SP and SPA (by over an order of magnitude). In general we feel that past literature has not adequately characterized the relative performance of the wide array of selectivity estimation methods that have be proposed over the years.

We do not claim incorrect the results reported in past literature. Rather, that those studies did not compare their methods across a wide enough array of datasets. For example, only one of the above studies reported results on real data (SMP, [15]) and they used only one dataset. The rest reported results only on synthetic data, but [10] (LH) claim their reported results to be comparable to that on real datasets not reported. We observed different behavior among real datasets and between them and synthetic data.

[8] They incorporated feedback in LH, but we did not in our experiments for reasons discussed earlier.

8 Conclusions

We introduced a novel type of histogram using information theory for one dimensional selectivity estimation. We evaluated the performance of this histogram empirically against a large number of known methods for selectivity estimation in the database literature and a large number of real datasets.

We observed the entropy histograms to fare well on real data. While they do not outperform all methods on all datasets, neither do any other methods. The entropy histograms outperformed all other methods on 4 out of 9 datasets and tied for first on another two. The closest competitors are the spline histograms which outperforms all others on 3 out of 9 datasets and tied on another one. The entropy histograms are an excellent choice of summary structure with respect to the state-of-the-art.

We also observed that all methods demonstrate a wide variety of behavior across real and synthetic datasets *i.e.* the accuracy rankings across datasets change substantially. For example, linear wavelets is tied for first on one real dataset and last on another. Also, MDA and OVA outperforms all other methods on a synthetic dataset, but rank nearly last on a real dataset. Moreover, we also observe results not consistent with many conclusions drawn in the literature.

One interesting general conclusion is that one-dimensional selectivity estimation is a complex problem involving many variables that have a significant impact. In order to understand well the behavior of a method, one must test it on a wide spectrum of datasets. We believe that the literature has not adequately characterized the performance of previous techniques.

References

[1] Blake C. and Merz C. UCI Repository of Machine Learning Databases. www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[2] Buccafurri F., Pontieri L., Rosaci D., and Sacca' D. Improving Range Query Estimation on Histograms. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pages 628–638, 2002.

[3] Cover T. and Thomas J. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[4] Garcia-Molina H., Ullman J., and Widom J. *Database System Implementation*. Prentice Hall, Upper Saddle River, NJ, 2000.

[5] Gilbert A., Kotidis Y., Muthukrishnan S., and Strauss M. Optimal and Approximate Computation of Summary Statistics for Range Aggregates. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principals of Database Systems (PODS)*, pages 227–236, 2001.

[6] Guha S., Koudas N., and Srivastava D. Fast Algorithms for Hierarchical Range Histogram Construction. In *Proceedings of the 21th ACM SIGMOD-SIGACT-SIGART Symposium on Principals of Database Systems (PODS)*, pages 180–187, 2002.

[7] Hettich S. and Bay S. UCI KDD Archive. University of California, Irvine, Department of Information and Computers Sciences, kdd.ics.uci.edu, 1999.

[8] Ioannidis Y. A History of Histograms. In *Proceedings of the 29th International Conference on Very Large Databases (VLDB)*, pages 19–30, 2003.

[9] Jagadish H., Koudas N., Muthukrishnan S., Poosala V., Sevcik K., and Suel T. Optimal Histograms with Quality Guarantees. In *Proceedings of the 24th International Conference on Very Large Databases (VLDB)*, pages 275–286, 1998.

[10] Konig A. and Weikum G. Combining Histograms and Parametric Curve Fitting for Feedback-Driven Query Result-Size Estimation. In *Proceedings of the 15th International Conference on Very Large Databases (VLDB)*, pages 423–434, 1999.

[11] Matias Y., Vitter J., and Wang M. Wavelet-Based Histograms for Selectivity Estimation. In *Proceedings of the 1998 International Conference on Management of Data (SIGMOD)*, pages 448–459, 1998.

[12] Poosala V., Ganti V., and Ioannidis Y. Approximate Query Answering using Histograms. *IEEE Data Engineering Bulletin*, 22(4):5–14, 1999.

[13] Poosala V., Ioannidis Y., Haas P., and Shekita E. Improved Histograms for Selectivity Estimation of Range Predicates. In *Proceedings of the 1996 International Conference on Management of Data (SIGMOD)*, pages 294–305, 1996.

[14] Sweldens W. and Schr¨oder P. Building your own wavelets at home. In *Wavelets in Computer Graphics*, pages 15–87. ACM SIGGRAPH Course notes, 1996.

[15] Wu L., Agrawal D., and El Abbadi A. Applying the Golden Rule of Sampling for Query Estimation. In *Proceedings of the 2001 International Conference on Management of Data (SIGMOD)*, pages 449–460, 2001.

[16] Zhang Q. and Lin X. On Linear-Spline Based Histograms. In *Lecture Notes in Computer Science 2419 (Proceedings of the 3rd International Conference on Web Age Information Management (WAIM'2002))*, pages 354–366, 2002.