ARTICLE IN PRESS

Data & Knowledge Engineering xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Improving estimation accuracy of aggregate queries on data cubes

E. Pourabbas [a,*], A. Shoshani [b,1]

[a] Italian National Research Council, Istituto di Analsi dei Sistemi ed Informatica "Antonio Ruberti", Viale Manzoni 30, 00185 Rome, Italy
[b] Lawrence Berkeley National Laboratory, Mailstop 50B-3238, 1 Cyclotron Road, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Query estimation
Entropy
Accuracy analysis

ABSTRACT

In this paper, we investigate the problem of estimation of a target database from summary databases derived from a base data cube. We show that such estimates can be derived by choosing a primary database with the desired target measure but not the desired dimensions, and use a proxy database to estimate the results. This technique is common in statistics, but an important issue we are addressing is the accuracy of these estimates. Specifically, given multiple primary and multiple proxy databases, the problem is how to select the primary and proxy databases that will generate the most accurate target database estimation possible. We propose an algorithmic approach which makes use of the principles of information entropy for determining the steps to select or compute the primary and proxy databases that provide the most accurate target database. We show that the primary database with the largest number of cells in common with the target database and the proxy database provides the more accurate estimates. We prove that this is consistent with maximizing the entropy. We provide some experimental results on the accuracy of the target database estimation in order to verify our results. Furthermore, we investigate the accuracy results in cases where the dimensions are defined over a hierarchy of categories and roll-up and drill-down operations are needed to generate the desired target results.

1. Introduction

Providing exact answers to queries from large data cubes in OLAP applications can be too slow, and in some cases, the user may prefer fast approximate answers. A more crucial case is when it is not possible to provide precise answers, such as in socio-economic applications because only summarized data is available for reasons of privacy. In such cases, it is quite useful to generate an estimate or approximate answers using approximate query processing techniques. A key issue is the accuracy of the estimates for aggregate queries (e.g., queries computing SUM or COUNT expressions), which was the focus of recent research activity (e.g., [3,14,15]).

In a previous paper [15], we discussed the estimation of summary queries, evaluated over multiple source summary databases. Such a summary query consists of requesting a summary measure of interest (e.g., household income), called *target measure*, over a set of category attributes, called *target dimensions* (e.g., *State*, *Sex*). In many cases, it may not be possible to evaluate such a query from a single source summary database, and two summary databases have to be used. For example, suppose that one database contains *Income by* (*State, Age, Race*) and the second contains *population by* (*State, Age, Education_level, Sex*). It is possible to estimate the target database *Income by* (*State, Sex*) by using the first database as the "primary"

* Corresponding author. Tel.: +39 06 7716481; fax: +39 06 7716461.
E-mail addresses: pourabbas@iasi.cnr.it, elaheh.pourabbas@iasi.cnr.it (E. Pourabbas).

[1] This work was supported by the Director, Office of Science, of the US Department of Energy under Contract No. DE-AC02-05CH11231.

0169-023X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2009.08.010

database (since it has the target measure *Income*), and using the second database as a "proxy" database (since it has the additional desired target dimension *Sex*). Here the population sizes are considered a proxy for the measure *Income*. The estimation method used to generate the target database is the linear indirect estimator (see Appendix A), which takes advantage of the fact that the summary databases were derived from the same base data, and consequently are correlated. The proposed method to estimate efficiently the target database was based on partitioning the dimensions of the source databases into three types: "target", "common", and "non-common" dimensions. We first determine the target dimensions, and classify the remaining dimensions as common and non-common. In the example above, *State* and *Sex* are target dimensions, *Age* is a common dimension, and *Race* and *Education_level* are non-common dimensions.

In the previous paper mentioned above, we examined two obvious computational methods for computing such a target database, called the "Full cross product" (F) and the "Pre-aggregation" (P) methods. Essentially, the estimation by the F method is achieved by first calculating the target measure over the full cross product of the dimensions from both databases using proportional estimation, and then aggregating over all the non-target dimensions. Since this method requires generating the full cross product, its cost is high. In contrast, the estimation by the P method consists of aggregating over all the non-target dimensions of both databases first, and only then generating the cross product using proportional estimation to obtain the result. The pre-aggregation reduces the size of the cross product greatly, and lowers the cost of generating the estimation. However, we showed that the P method, while computationally efficient, yields results that are not as accurate as the F method. We proposed a third method called "Partial Pre-aggregation" (PP) method, which consists of summarizing only the non-common dimensions first, and then applying the proportional estimation. Using a measure of accuracy, called average relative error (ARE) (see Appendix B), we proved that the PP method yields the same accuracy as the F method, but reduces significantly the computational and space complexity. The reduction in cost is by a factor proportional to the multiplication of the cardinalities of the non-common dimensions.

In this paper, we consider an open question which was left as future challenge in [15]. The question is how to select a primary and a proxy database given that there are multiple primary databases available with the same measure and multiple proxy databases with the desired target dimensions in order to get the most accurate estimation results. This paper is an extension version of our paper published on the proceedings of DOLAP 2008 [16]. In this extended version, we perform additional experiments and investigate the accuracy results in cases where the dimensions are defined over a hierarchy of categories and roll-up and drill-down operations are needed to generate the desired target results.

1.1. The problem

To explain the idea let us consider the following multiple primary databases:

- $DB_{PR1} = Income(State, Age)$
- $DB_{PR2} = Income(State, Labor\_status)$
- $DB_{PR3} = Income(Age, Labor\_status)$
- $DB_{PR4} = Income(State, Age, Labor\_status)$

and multiple proxy databases:

- $DB_{PX1} = Population(State, Age, Sex)$
- $DB_{PX2} = Population(State, Labor\_status, Sex)$
- $DB_{PX3} = Population(Age, Labor\_status, Sex)$
- $DB_{PX4} = Population(State, Age, Labor\_status, Sex)$

where the cardinalities of the dimensions are: $|\mathrm{State}| = 52, |\mathrm{Age}| = 4, |\mathrm{Labor\_status}| = 2$, and $|\mathrm{Sex}| = 2$. Note that the two categories of *Labor_status* are *In_ Labor_Force* and *Not_in_Labor_Force* according to US Census Bureau.[2] Let *Income(State, Age, Labor_status, Sex)* be the target database, which should be estimated from the sets of summary databases given above. If we select the first primary database, i.e., *Income(State, Age)*, then we can apply DB_{PX2}, DB_{PX3}, and DB_{PX4} to estimate the target database since only these proxy databases contain auxiliary data on the dimensions *Labor_status* and *Sex*. Similarly, if we choose the second primary database, we can only apply DB_{PX1}, DB_{PX3}, and DB_{PX4}. The third primary database needs auxiliary data on dimensions *State* and *Sex*, which are provided by DB_{PX1}, DB_{PX2}, and DB_{PX4}. Whereas, for the last primary database all four proxy databases can be applied. This is labeled as *Case 1* in Table 1, where we assume that all four primary databases exist, as well as all four proxy databases exist. We also include in Table 1 three additional cases where only some of the primary or proxy databases are shown. These cases will be used later to illustrate situations that require special attention.

In all four cases, as we mentioned before, the main goal is to obtain more accurate estimated results for the target database. Thus, to achieve this goal we have to select two source databases, one primary and one proxy databases. The problem is which databases we should choose from a given set of primary and proxy databases that provide the more accurate estimation results.

[2] http://www.census.gov.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

Table 1
Cases.

Cases	Primary DBs	Proxy DBs
Case (1)	DB_{PR1}	DB_{PX1}
	DB_{PR2}	DB_{PX2}
	DB_{PR3}	DB_{PX3}
	DB_{PR4}	DB_{PX4}
Case (2)	DB_{PR1}	DB_{PX1}
	DB_{PR2}	DB_{PX2}
	DB_{PR3}	DB_{PX3}
		DB_{PX4}
Case (3)	DB_{PR1}	DB_{PX1}
	DB_{PR2}	DB_{PX2}
	DB_{PR3}	DB_{PX3}
	DB_{PR4}	
Case (4)	DB_{PR1}	DB_{PX1}
	DB_{PR2}	DB_{PX2}
	DB_{PR3}	DB_{PX3}

The solution of the problem mentioned above is based on two conjectures. The first one is that the more cells of common dimensions the primary database shares with the target database, the more accurate are the estimated results. A cell is defined as the smallest element formed by the cross product of the dimensions. Referring to the primary databases shown in *Case 1*, DB_{PR4} not only shares the largest number of cells of common dimensions with the target database but also covers all the dimensions of the first three primary databases. Note that in this case all common dimensions are target dimensions. Now, let us consider *Case 2* and *Case 4*. The problem is which primary database should we choose? In the next section, we will show that basing this decision on the estimate of the maximum entropy provides the most accurate results possible.

The second conjecture is that the proxy database that shares the largest number of cells of the common dimensions with the primary database provides more accurate results. In *Case 1* and *Case 2*, DB_{PX4} is such a proxy database. A similar problem arises when selecting the proxy database in *Case 3* and *Case 4*. In these cases, which approach should be applied in order to select the proxy database for the estimation of the target database? We discuss this problem in the next section as well.

1.2. Related work

There was a significant amount of work in the literature on approximate query processing. In [11], for instance, the definition of a universal statistical database containing several summary tables which share the same summary measure is examined. Given a query, a system of linear equations over the universal database is constructed whose solutions satisfy the query. In [12,13], the problem of evaluating a summary query from a set of summary tables sharing the same variable and an auxiliary table is discussed. These works propose algorithms which make use of techniques developed in the theory of acyclic database schemas. In contrast, we focus here on the problem of the accuracy of the query estimation. In our work, we consider a set of proxy (or auxiliary) databases, which share the same summary measures.

In [5] the authors propose a framework for approximate answers to aggregation queries called online aggregation in which the base data is scanned in random order at query time and the approximate answer is continuously updated as the scan proceeds. The approximate query answering (AQUA) [1] system provides approximate answers using small pre-computed synopses of the underlying base data. In [14], the authors consider the problem of deriving approximately the original data from the aggregates. They propose a framework for estimating the original values based on the notion of information entropy. In our work, we use a different approach of estimating the values of the target database by using additional information from proxy databases. We apply the principles of entropy over the multiple source databases in order to identify two databases, one primary database and one proxy database, which achieve more accurate results. We prove formally that the source databases with the largest number of cells in common provide the most accurate results possible. Based on these results, we propose an algorithmic approach for determining the steps to select or compute the source databases from multiple summary databases.

The paper is structured as follows. The next section provides the principles of entropy used in this paper. In this section we also introduce the formal model which provides the basis for a formal analysis of the results in this paper. Section 3 discusses the problem of selecting two source summary databases from multiple primary and multiple proxy databases in order to achieve maximum accuracy for the target database. In Section 4, we develop an algorithmic approach for determining the steps to achieve maximum accuracy, and we prove theorems which show that the source databases with the largest number of cells in common provide the more accurate estimates. Section 5 illustrates some experimental results on the accuracy of the target database estimation. In Section 6 we consider the accuracy of results in cases where dimensions have a hierarchical structure to them, and roll-up or drill-down operations are needed in order to generate the desired target results. Section 7 contains the conclusions.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

2. Principles and formal model

2.1. Principles of entropy

In this section, we recall the principles of maximum entropy and minimum cross-entropy, which will be used in the next sections. The (Shannon) *entropy* H of a discrete probability distribution $p(x)$ is the non-negative function

$$H(p) = -\sum_{x \in X} p(x) log\ p(x) \tag{1}$$

where X represents the set of instances. H reaches its maximum value at the uniform distribution over X, i.e., $log|X|$. In statistics and information theory, a maximum entropy probability distribution is a probability distribution whose entropy is at least as great as that of all other members of a specified class of distributions.

Let $P(X_1, \ldots, X_n)$ be an n-dimensional discrete probability distribution to be estimated from $P'(X_1, \ldots, X_n)$ and the set of all marginal distribution $P_i(X_i)$ with $i = 1, \ldots, n$ ("Marginals" is a commonly-used term in Statistics that refers to the summary of rows and columns in the "margins" of a table). If $X = \{X_1, \ldots, X_n\}$, we may find P that maximizes the entropy $H(P)$ of P over all marginal probability distributions such that it satisfies the following constrains:

- every element in $P(X)$ is non-negative value
- $\sum P(X) = 1$
- $P(X_i) = P_i(X_i)$

Note that in this paper, we will refer to the constraints mentioned above as the *consistency conditions*. Let $\widehat{P}(X)$ be the maximum entropy approximation to $P(X)$. The *cross-entropy* (or *relative entropy* or *Kullback–Leibler distance*) between $\widehat{P}(X)$ and $P(X)$ measures the similarity of two distribution and is defined as follows:

$$D(\widehat{P}, P) = \sum_{X} \widehat{P}(X) log \frac{\widehat{P}(X)}{P(X)} \tag{2}$$

Minimizing $D(\widehat{P}, P)$ is the same as maximizing the entropy of P. The technique used to compute the maximum entropy estimate is *Iterative Proportional Fitting Procedure*-IPFP [4], which starts with the *zero approximation* $P^{[0]}(X) = P'(X)$ and determines the *higher-order approximations* to $P(X)$ according to the following computation scheme:

first iteration cycle	$P^{[1]}(X)$	...	$P^{[n]}(X)$
second iteration cycle	$P^{[n+1]}(X)$	...	$P^{[2n]}(X)$
...			
h-th *iteration cycle*	$P^{[hn+1]}(X)$	...	$P^{[hn+n]}(X)$
...			

where the approximation $P^{[hn+i]}(X)$ in the $(h+1)$-th iteration cycle, $1 \leqslant i \leqslant n$, is obtained by fitting the approximation $P^{[hn+i-1]}(X)$ to the marginal distribution $P_i(X_i)$ as follows:

$$P^{[hn+i]}(X) = \frac{P_i(X_i)}{P^{[hn+i-1]}(X_i)} P^{[hn+i-1]}(X).$$

This procedure converges monotonically to the maximum entropy estimation. The iterations stop when the estimate at two consecutive steps are the same or the difference of estimates are less than a pre-defined value.

2.2. Formal model

We use here the formal model defined in [15], which provides the basis for a formal analysis of the results. In the following sections, we assume two source summary databases, called DB_P and DB_Q that are used to produce a *target database* DB_T. The databases are defined as follows: $DB_P = M_P\left(\left\{A_P^i 0 < i \leqslant m\right\}\right)$, $DB_Q = M_Q\left(\left\{A_Q^j 0 < j \leqslant n\right\}\right)$, and $DB_T = M_T\left(\left\{A_T^k 0 < k \leqslant t\right\}\right)$, where M_P, M_Q, and M_T are the measures of the corresponding databases, A_P^i, A_Q^j, and A_T^k are the corresponding dimensions, and m, n, and t are the cardinalities of the corresponding dimensions. In defining a target database over the two source summary databases, one of the measures, either M_P or M_Q is selected. Without loss of generality, suppose that M_P is selected. Thus, $M_P = M_T$. DB_P is called the *primary database*, M_Q is called the *proxy measure*, and DB_Q is called the *proxy database*.

Given two source summary databases DB_P and DB_Q that are used to generate a target database DB_T, we can classify the source database dimensions as belonging to three disjoint groups: target dimensions, common dimensions, and non-common dimensions. First, we pick the dimensions in the source databases that are specified in the target database for the target group; then the *remaining* dimensions are considered common if they are in both source databases, and are considered non-common otherwise. Note that a target dimension can exist in both source databases. We use the following notation:

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$DB_P = M_P\left(A_P^C, A_P^{\overline{C}}, A_P^{T^C}, A_P^{T^{\overline{C}}}\right)$, and $DB_Q = M_Q\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^C}, A_Q^{T^{\overline{C}}}\right)$, where $C, \overline{C}$, and T refer to the common, non-common, and target dimension-groups, respectively. Note that $A_P^C = A_Q^C$, and $A_P^{T^C} = A_Q^{T^C}$. We use the notation A_T for the group of target dimensions $\left\{A_T^k 0 < k \leqslant t\right\}$. Thus, $DB_T = M_T(A_T)$. Using the notation above, we have $A_T = A_P^{T^C} \cup A_P^{T^{\overline{C}}} \cup A_Q^{T^{\overline{C}}}$. Note that $A_Q^{T^{\overline{C}}}$ must always exist to make the proxy summarization meaningful. However, $A_P^{T^C}$ and $A_P^{T^{\overline{C}}}$ may or may not exist. Indeed, if $A_Q^{T^{\overline{C}}}$ does not exist, then there is no need to use DB_Q, since the results can be obtained from DB_P only.

For instance, let us consider the source summary databases: *Income(Age, Labor_status, Sex)*, and *Population(State, Age, Race, Sex)*. Let us assume that the summary query expressed over them is *Income(State)*. In this case, *Income(State)* is the target summary database, *Population(State, Age, Race, Sex)* is the proxy database, and *Income(Age, Labor_status, Sex)* is the primary database. $A_T = \{State\}$ is the target dimension, where $A^{T^C}_{Population} = A^{T^C}_{Income} = \varnothing, A^{T^{\overline{C}}}_{Population} = \{State\}, A^{T^{\overline{C}}}_{Income} = \varnothing$ are the non-common target dimensions, $A^{C}_{Population} = A^{C}_{Income} = \{Age, Sex\}$ are the common dimensions between the source summary databases, and $A^{\overline{C}}_{Population} = \{Race\}$, and $A^{\overline{C}}_{Income} = \{Labor\_status\}$ are the non-common dimensions. As another example, consider the case where the summary query expressed over the source databases is *Income(State, Age)*, then $A_T = \{State, Age\}$ and accordingly, $A^{T^C}_{Population} = A^{T^C}_{Income} = \{Age\}, A^{T^{\overline{C}}}_{Population} = \{State\}, A^{T^{\overline{C}}}_{Income} = \varnothing$, and $A^{C}_{Population} = A^{C}_{Income} = \{Sex\}$.

3. Database selection

In this section, we investigate the problem of selecting two source summary databases from multiple primary and multiple proxy databases in order to achieve maximum accuracy for the target database. Only primary databases that have the same measure as that of the target database need be considered.

The proxy database is selected in order to provide the dimensions missing in the primary database and specified in the target database. For all four cases shown in Section 1.1, the *Sex* dimension in the proxy databases is needed for the target database and is not available from primary databases. We recall the results discussed in [15] regarding the non-common dimensions or the dimensions which are not specified in the target database but exist in one of the source databases. According to the Partial Pre-aggregation (PP) method, pre-aggregating the source databases over the non-common dimensions, the estimation results are as accurate as the estimates obtained by first generating the full cross product of all dimensions of the source databases and then aggregating over non-common dimensions. In this paper, we use this approach in considering which primary and proxy databases to choose to maximize accuracy.

In the previous section, we conjectured that the primary database which includes the largest number of cells of the desired target dimensions is the better choice. Let us recall the set of primary databases shown in *Case 1*, and shown in Table 2 (where we use the symbol "*I*" to indicate *Income*.) By multiplying the cardinalities of the dimensions we obtain the number of cells for each choice. We use the notation $|A|$ in Table 2 for this product of cardinalities. As can be seen in Table 2, DB_{PR4} shares 416 cells for dimensions in common with the target database $Income(State, Age, Labor\_status, Sex)$. It includes more cells with respect to the other three primary databases. An important idea associated with the number of cells is that of entropy. According to the principles discussed in Section 2.1, given a set of primary databases we have to choose the one with the largest number of cells to achieve the largest entropy [7]. In Section 4 we prove in the first theorem that the more accurate estimate is achieved when the primary database with the largest number of cells in common with the target database is selected. For the databases shown in Table 2, the largest entropy is achieved by DB_{PR4}. This primary database also satisfies the three constraints of consistency conditions listed in Section 2.1. Concerning the proxy databases (see Table 3 where the symbol "*P*" refers to *Population*), if there are common dimensions, we conjecture that the proxy database with the largest number of cells of the common dimensions with the primary database achieves the more accurate result. In this case, it is DB_{PX4}. This conjecture is also proven in Section 4 where we show in the second theorem that the more accurate estimate is achieved when the proxy database with the largest number of cells in common with the primary database is selected.

The relative entropy (or loss of information) of the estimates by applying each primary database to DB_{PX4} is shown in Table 2, fourth column. As can be seen, for DB_{PR4}, the amount of information that we lose is less than the others. This indicates that the estimate obtained by DB_{PR4} is more similar to that of the real distribution of *Income* with respect to the other primary databases. Thus, the combination of DB_{PR4} and DB_{PX4} provides the more accurate estimate. The accuracy results are given in Section 5.

Table 2
Primary databases.

PrimaryDB	$\lvert A\rvert$	*Entropy*	$D(\hat{I}, I)$
$DB_{PR1} = I(State, Age)$	208	6.45	0.06816
$DB_{PR2} = I(State, Labor\_status)$	104	5.54	0.09071
$DB_{PR3} = I(Age, Labor\_status)$	8	3.49	0.13815
$DB_{PR4} = I(State, Age, Labor\_status)$	416	7.10	0.01623

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

Table 3
Proxy databases.

Proxy DB	$\|A\|$
$DB_{PX1} = P(State, Age, Sex)$	416
$DB_{PX2} = P(State, Labor\_status, Sex)$	208
$DB_{PX3} = P(Age, Labor\_status, Sex)$	16
$DB_{PX4} = P(State, Age, Labor\_status, Sex)$	832

Now, suppose in Table 2 that only the first three databases are given (i.e., *Case 2*). In this case, the maximum number of cells is provided by DB_{PR1}, but none of them satisfies the consistency conditions (see Section 2.1). Thus, $Income(State, Age, Labor\_status)$ needs to be estimated. For this reason, we have to consider all three primary databases by applying IPFP to estimate $\widehat{I}ncome(State, Age, Labor\_status)$. This estimate satisfies the above mentioned condition because, for instance, aggregating that over "Age", we have $Income(State, Labor\_status)$, over "Labor_status" we obtain $Income(State, Age)$ and over "State" we obtain $Income(Age, Labor\_status)$. This estimate provides maximum entropy and contains the largest number of cells in common with the target database (this is expressed in the PROCEDURE in Section 4). In [13], it is discussed that this estimate is uniquely determined by the information-theoretic principle of *minimum cross-entropy* and its distribution is defined as follows. (For the sake of brevity, the symbols "*S*", "*A*" , and "*L*" indicate "*State*", "*Age*", and "*Labor_status*", respectively.)

$$\widehat{I}ncome[0](S,A,L) = P(S,A,L)$$
$$\widehat{I}ncome[1](S,A,L) = Income(S,A)\frac{\widehat{I}ncome[0](S,A,L)}{\sum_L \widehat{I}ncome[0](S,A,L)}$$
$$\widehat{I}ncome[2](S,A,L) = Income(S,L)\frac{\widehat{I}ncome[1](S,A,L)}{\sum_A \widehat{I}ncome[1](S,A,L)}$$
$$\widehat{I}ncome[3](S,A,L) = Income(A,L)\frac{\widehat{I}ncome[2](S,A,L)}{\sum_S \widehat{I}ncome[2](S,A,L)}$$
$$\widehat{I}ncome[4](S,A,L) = Income(S,A)\frac{\widehat{I}ncome[3](S,A,L)}{\sum_L \widehat{I}ncome[3](S,A,L)}$$
$$\ldots$$

Note that the zero approximation (or initial distribution) is set to the proxy database with the same dimensions of the estimate of *Income*. In this example, the proxy is DB_{PX4}, where $P(S,A,L) = \sum_{Sex} P(S,A,L,Sex)$.

Case 4 differs from *Case 2* in the proxy database computation. In order to apply IPFP to the primary databases, the zero approximation should be set to $P(S,A,L)$, but this proxy is not provided. Our solution is to estimate $\widehat{P}(S,A,L,Sex)$ from the proxy databases. We return to this point in Section 5. The estimate of the primary database is obtained by IPFP, where the zero approximation is defined by the aggregation over *Sex* of $\widehat{P}(State, Age, Labor\_status, Sex)$ given below:

$$\widehat{P}(State, Age, Labor\_status, Sex) = P(State, Age, Sex)\frac{P(State, Labor\_status, Sex)}{\sum_{Labor\_status} P(State, Labor\_status, Sex)}$$

As a final remark, we emphasize that in each set of databases there can be summary databases which are marginal of a database in the same set. They are not considered in the database selection because they are redundant.

4. Algorithmic approach

We propose the use of an algorithmic approach for determining the steps to achieve maximum accuracy. The procedure is essentially based on two theorems introduced below. Using the notation introduced in Section 2.2, we can formulate the following definition and theorems.

Definition 4.1. Let $M_{P_k}\left(A^C_{P_k}, A^{\overline{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}\right), M_{P_l}\left(A^C_{P_l}, A^{\overline{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}\right)$ be primary databases, and let $M_Q\left(A^C_Q, A^{\overline{C}}_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q\right)$ be a proxy database. We define $\widehat{M}_{P_k}$ to be the estimation result of the target database over the primary summary database $M_{P_k}\left(A^C_{P_k}, A^{\overline{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}\right)$. Similarly, we define $\widehat{M}_{P_l}$ to be the estimation result of target database over the primary database $M_{P_l}\left(A^C_{P_l}, A^{\overline{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}\right)$. The expressions of the estimators above are defined by applying the PP method, according to which the source databases are aggregated over non-common dimensions first:

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right) = \sum_{A^{\bar{C}}_{P_k}} M_{P_k}\left(A^C_{P_k}, A^{\bar{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)$$

$$M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right) = \sum_{A^{\bar{C}}_{P_l}} M_{P_l}\left(A^C_{P_l}, A^{\bar{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)$$

$$M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right) = \sum_{A^{\bar{C}}_Q} M_Q\left(A^C_Q, A^{\bar{C}}_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)$$

then, linear indirect estimation method is applied:

$$\widehat{M}_{P_k}\left(A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}, A^C_Q, A^{T^{\bar{C}}}_Q\right) = M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_Q, A^{T^C}_Q\right)}$$

$$\widehat{M}_{P_l}\left(A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}, A^C_Q, A^{T^{\bar{C}}}_Q\right) = M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_Q, A^{T^C}_Q\right)}$$

where, $M_Q\left(A^C_Q, A^{T^C}_Q\right) = \sum_{A^{T^{\bar{C}}}_Q} M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)$.

Theorem 4.1. *Let* $M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right), M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)$ *be primary databases, and let* $M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)$ *be a proxy database, where* $|A_{P_l}| < |A_{P_k}| < |A_T|, A^{\mathscr{C}}_{P_l} \subset A^{\mathscr{C}}_{P_k}$, *and* $\mathscr{C}$ *represents common and common-target dimension-groups. Let* $\widehat{M}_{P_k}$ *and* $\widehat{M}_{P_l}$ *be the estimate of the target database obtained by applying the primary databases* M_{P_k} *and* M_{P_l} *to* M_Q, *respectively. The primary database* M_{P_k} *achieves better estimates with respect to* M_{P_l}.

Proof. Let the relative entropy of $\widehat{M}_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}'}}_{P_k}, A^{T^{\bar{C}}}_Q\right)$ and $\widehat{M}_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}'}}_{P_l}, A^{T^{\bar{C}}}_Q\right)$ with respect to $M_P\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}, A^{T^{\bar{C}}}_Q\right)$ be defined according to expressions:

$$D(\widehat{M}_{P_k}, M_P) = \sum \widehat{M}_{P_k} log \frac{\widehat{M}_{P_k}}{M_P}$$

$$= \sum \left(\left(M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_k}, A^{T^C}_{P_k}\right)} \right) log \frac{M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_k}, A^{T^C}_{P_k}\right)}}{M_P\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}, A^{T^{\bar{C}}}_Q\right)} \right)$$

$$D(\widehat{M}_{P_l}, M_P) = \sum \widehat{M}_{P_l} log \frac{\widehat{M}_{P_l}}{M_P}$$

$$= \sum \left(\left(M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_l}, A^{T^C}_{P_l}\right)} \right) log \frac{M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_l}, A^{T^C}_{P_l}\right)}}{M_P\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}, A^{T^{\bar{C}}}_Q\right)} \right)$$

We show $D(\widehat{M}_{P_k}, M_P) < D(\widehat{M}_{P_l}, M_P)$, or $D(\widehat{M}_{P_l}, M_P) - D(\widehat{M}_{P_k}, M_P) > 0$ as follows:

$$D(\widehat{M}_{P_l}, M_P) - D(\widehat{M}_{P_k}, M_P) = \sum \left(\left(M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_l}, A^{T^C}_{P_l}\right)} \right) log \frac{M_{P_l}\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_l}, A^{T^C}_{P_l}\right)}}{M_P\left(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\bar{C}}}_{P_l}, A^{T^{\bar{C}}}_Q\right)} \right)$$

$$- \sum \left(\left(M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_k}, A^{T^C}_{P_k}\right)} \right) log \frac{M_{P_k}\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}\right)\frac{M_Q\left(A^C_Q, A^{T^C}_Q, A^{T^{\bar{C}}}_Q\right)}{M_Q\left(A^C_{P_k}, A^{T^C}_{P_k}\right)}}{M_P\left(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\bar{C}}}_{P_k}, A^{T^{\bar{C}}}_Q\right)} \right)$$

Setting

$$\mathscr{F} = M_Q\left(A_Q^C, A_Q^{T^C}, A_Q^{T^{\bar{C}}}\right)\frac{M_{P_k}\left(A_{P_k}^C, A_{P_k}^{T^C}, A_{P_k}^{T^{\bar{C}}}\right)M_Q\left(A_{P_l}^C, A_{P_l}^{T^C}\right)}{M_Q\left(A_{P_k}^C, A_{P_k}^{T^C}\right)M_{P_l}\left(A_{P_l}^C, A_{P_l}^{T^C}, A_{P_l}^{T^{\bar{C}}}\right)}, \quad \mathscr{G} = \frac{M_{P_l}\left(A_{P_l}^C, A_{P_l}^{T^C}, A_{P_l}^{T^{\bar{C}}}\right)M_Q\left(A_{P_k}^C, A_{P_k}^{T^C}\right)}{M_Q\left(A_{P_l}^C, A_{P_l}^{T^C}\right)M_{P_k}\left(A_{P_k}^C, A_{P_k}^{T^C}, A_{P_k}^{T^{\bar{C}}}\right)},$$

$$D\left(\widehat{M}_{P_l}, M_P\right) - D\left(\widehat{M}_{P_k}, M_P\right) = \sum \mathscr{F}\mathscr{G}\ log\ \frac{M_{P_l}\left(A_{P_l}^C, A_{P_l}^{T^C}, A_{P_l}^{T^{\bar{C}}}\right)M_Q\left(A_{P_k}^C, A_{P_k}^{T^C}\right)}{M_Q\left(A_{P_l}^C, A_{P_l}^{T^C}\right)M_{P_k}\left(A_{P_k}^C, A_{P_k}^{T^C}, A_{P_k}^{T^{\bar{C}}}\right)} = \sum \mathscr{F}\mathscr{G}\ log\ \mathscr{G}$$

Since $\sum \mathscr{F}\mathscr{G} = \sum M_Q\left(A_Q^C, A_Q^{T^C}, A_Q^{T^{\bar{C}}}\right) = 1$, and according to Theorem 3.1 (The theorem states the relative entropy obtained from distributions of the observations is positive, see Chapter 2) in [8]

$$\sum \mathscr{G}\ log\ \mathscr{G} = D\left(M_{P_l}\left(A_{P_l}^C, A_{P_l}^{T^C}, A_{P_l}^{T^{\bar{C}}}\right), \quad M_Q\left(A_{P_l}^C, A_{P_l}^{T^C}\right)\right) + D\left(M_{P_k}\left(A_{P_k}^C, A_{P_k}^{T^C}, A_{P_k}^{T^{\bar{C}}}\right), \quad M_Q\left(A_{P_k}^C, A_{P_k}^{T^C}\right)\right)$$

which leads to the conclusion that $\sum \mathscr{F}\mathscr{G}\ log\ \mathscr{G} > 0$. Thus, $D(\widehat{M}_{P_l}, M_P) - D(\widehat{M}_{P_k}, M_P) > 0$, with equality if and only if:

$$\frac{M_{P_l}\left(A_{P_l}^C, A_{P_l}^{T^C}, A_{P_l}^{T^{\bar{C}}}\right)}{M_Q\left(A_{P_l}^C, A_{P_l}^{T^C}\right)} = \frac{M_{P_k}\left(A_{P_k}^C, A_{P_k}^{T^C}, A_{P_k}^{T^{\bar{C}}}\right)}{M_Q\left(A_{P_k}^C, A_{P_k}^{T^C}\right)} \qquad \square$$

Definition 4.2. Let $M_P\left(A_P^C, A_P^{\bar{C}}, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)$, be primary database, and let $M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{\bar{C}}, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right), M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{\bar{C}}, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)$ be proxy databases. We define $\widehat{M}_{P_k}$ to be the estimation result of the target database by applying the primary database to $M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{\bar{C}}, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)$. Similarly, we define $\widehat{M}_{P_l}$ to be the estimation result of target database by applying the primary database to $M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{\bar{C}}, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)$. The expressions of the estimators above are defined by applying the PP method as follows:

$$\widehat{M}_{P_k}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_k}^C, A_{Q_k}^{T^{\bar{C}}}\right) = M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right)}$$

$$\widehat{M}_{P_l}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_l}^C, A_{Q_l}^{T^{\bar{C}}}\right) = M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)}$$

Theorem 4.2. *Let $M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)$, be primary database, and let $M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right), M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)$ be proxy databases, where $|A_{Q_l}| < |A_{Q_k}|, A_{Q_l}^{\mathscr{C}} \subset A_{Q_k}^{\mathscr{C}}$. Let $\widehat{M}_{P_k}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_k}^C, A_{Q_k}^{T^{\bar{C}}}\right)$ and $\widehat{M}_{P_l}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_l}^C, A_{Q_l}^{T^{\bar{C}}}\right)$ be the estimate of the target database obtained by applying the primary database $M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)$ to M_{Q_k} and M_{Q_l}, respectively. The estimate $\widehat{M}_{P_k}$ is more accurate than the estimate $\widehat{M}_{P_l}$.*

Proof. Let the relative entropy of $\widehat{M}_{P_k}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_k}^C, A_{Q_k}^{T^{\bar{C}}}\right)$ and $\widehat{M}_{P_l}\left(A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_l}^C, A_{Q_l}^{T^{\bar{C}}}\right)$ with respect to $M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}, A_Q^{T^{\bar{C}}}\right)$ be defined according to the following expressions:

$$D\left(\widehat{M}_{P_k}, M_P\right) = \sum \widehat{M}_{P_k} log \frac{\widehat{M}_{P_k}}{M_P}$$

$$= \sum \left(\left(M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right)}\right) log \frac{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right)}}{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_k}^{T^{\bar{C}}}\right)}\right)$$

$$D\left(\widehat{M}_{P_l}, M_P\right) = \sum \widehat{M}_{P_l} log \frac{\widehat{M}_{P_l}}{M_P}$$

$$= \sum \left(\left(M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)}\right) log \frac{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right)\frac{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)}}{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_l}^{T^{\bar{C}}}\right)}\right)$$

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

We show $D\left(\widehat{M}_{P_l}, M_P\right) - D\left(\widehat{M}_{P_k}, M_P\right) > 0$ as follows:

$$D\left(\widehat{M}_{P_l}, M_P\right) - D\left(\widehat{M}_{P_k}, M_P\right) = \sum \left(\left(M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right) \frac{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)} \right) log \frac{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right) \frac{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)}}{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_l}^{T^{\bar{C}}}\right)} \right)$$

$$- \sum \left(\left(M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right) \frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right)} \right) log \frac{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right) \frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right)}}{M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}, A_{Q_k}^{T^{\bar{C}}}\right)} \right)$$

Setting

$$\mathscr{F} = M_P\left(A_P^C, A_P^{T^C}, A_P^{T^{\bar{C}}}\right) \frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)},$$

$$\mathscr{G} = \frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)},$$

$$D\left(\widehat{M}_{P_l}, M_P\right) - D\left(\widehat{M}_{P_k}, M_P\right) = \sum \mathscr{F}\mathscr{G}\ log \frac{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\bar{C}}}\right)}{M_{Q_k}\left(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\bar{C}}}\right) M_{Q_l}\left(A_{Q_l}^C, A_{Q_l}^{T^C}\right)} = \sum \mathscr{F}\mathscr{G}\ log\ \mathscr{G}$$

Similar to Theorem 4.1, $\sum \mathscr{F}\mathscr{G}\ log\ \mathscr{G} > 0$ is shown.

To summarize the discussion above, the procedure for determining the steps to achieve maximum accuracy can be defined by PROCEDURE. It is composed of three parts. Note that in step (3), the second part is called for the propose of obtaining the proxy database which includes maximum common dimensions with the primary databases. □

5. Experimental results

PROCEDURE
Input: Given target database DB_T, multiple primary databases DB_{PRi} with $1 \leqslant i \leqslant n$ and multiple proxy databases DB_{PXj} $1 \leqslant j \leqslant m$ databases
Goal: Select two source databases to obtain maximum accuracy for the estimate of DB_T
PART 1 – SELECTION OF THE PRIMARY DATABASE
(1) Given that $M_T = M_{PR}$ start with selecting a primary database;
(2) Select the primary database whose dimensions cover the dimensions of all other primary databases (indicated by A_{PR})
(3) If no such primary database exists run PART 2 and then apply IPFP to multiple primary databases with zero approximation fixed to DB_{PX} pre-aggregated over $A_{PX}^{T^{\bar{C}}}$;
(4) Once DB_{PR} was chosen (step 2) or estimated (step 3), pre-aggregate the non-common dimensions;
PART 2 – SELECTION OF THE PROXY DATABASE
(5) Consider only DB_{PX} with dimensions $A_{PX} = A_{PX}^{T^{\bar{C}}} \cup A_{PR}$;
(6) If there is no such proxy database, consider proxy databases that have $A_{PX} = A_{PX}^{T^{\bar{C}}}$, and apply IPFP to multiple proxy databases;
(7) Once DB_{PX} was chosen (step 5) or estimated (step 6), pre-aggregate the non-common dimensions;
PART 3 – ESTIMATION OF THE TARGET DATABASE
(8) Apply linear indirect estimation method to DB_{PR}, and DB_{PX}.

We discuss the experimental results of the application of our algorithmic approach to the four cases introduced in Section 1.1. For the experimental results, we use the values in the base data to evaluate the estimated errors. We start with *Case 1*. We note that DB_{PR4} and DB_{PX4} satisfy step (2) and step (5). In fact, they provide the most accurate results (see Table 4, first

Table 4
Accuracy results of selected primary and proxy databases in four cases.

Cases	DB_{PR}	DB_{PX}	*ARE*
Case (1)	DB_{PR4}	DB_{PX4}	0.0962
Case (2)	$\widehat{I}(S,A,L)$	DB_{PX4}	0.1464
Case (3)	DB_{PR4}	$\widehat{P}(S,A,L,Sex)$	0.1186
Case (4)	$\widehat{I}(S,A,L)$	$\widehat{P}(S,A,L,Sex)$	0.1625

row). In *Case 2*, according to step (3), IPFP is applied to the given primary databases. As we mentioned in Section 3, the zero approximation is fixed to DB_{PX4} which is pre-aggregated over the non-common target dimension. The convergence of the estimate of Income is achieved after five iteration cycles. Note that, we could have fixed the zero approximation of IPFP to every primary database in order to estimate the primary database, but these starting values effect the accuracy of the results. In fact, the average relative error of the target database is 0.1732 vs. 0.1625 by applying step (3). Overall, the accuracy result in *Case 2* is close to that of *Case 4*. Similarly, the accuracy result in *Case 1* is close to that of *Case 3*. With respect to *Case 1*, the accuracy of *Case 3* is better than *Case 2*. It seems that the choice of the primary database effects the accuracy results more than the choice of the proxy database (see the accuracy of *Case 1* and *Case 2*). In contrast, the accuracy of *Case 4* is worse than the other cases.

Now, we compare accuracy results of the estimates. Specifically, in Table 5, we compare the accuracy results of the estimate of the target database by applying each primary database to $P(State, Labor\_status, Age, Sex)$ and to the estimate of the primary database computed according to step (3) of the proposed procedure. Table 6 illustrates the accuracy results of the estimate of the target database by applying to $I(State, Labor\_status, Age)$ each given proxy database and the estimated proxy database computed according to step (6) of the proposed procedure.

Finally, Table 7 shows the accuracy results of the estimate of the target database by applying the estimated primary database $\widehat{I}(State, Labor\_status, Age)$ to each given proxy database and the estimated proxy database $\widehat{P}(State, Labor\_status, Age, Sex)$. As can be seen, in all cases, when the consistency conditions do not hold, using the estimated databases generates the most accurate results.

5.1. IPFP and the F and PP methods

In this section, we investigate whether applying the PP method is useful for the IPFP procedure as well. That is, we investigate whether applying the PP method before performing the IPFP procedure, produces the same ARE as the F method. For a better illustration of the application of the PP method to the IPFP procedure, additional experiments were carried out. We

Table 5
ARE of $\widehat{I}(State, Labor\_status, Age, Sex)$ by applying the primary databases to $P(State, Labor\_status, Age, Sex)$.

Primary DB	$\|A\|$	*ARE*
$I(State, Age)$	208	0.3925
$I(State, Labor\_status)$	104	0.3991
$I(Age, Labor\_status)$	8	0.5300
$\widehat{I}(State, Age, Labor\_status)$	416	0.1464

Table 6
ARE of $\widehat{I}(State, Labor\_status, Age, Sex)$ by applying the primary database $I(State, Labor\_status, Age)$ to the following proxy databases.

Proxy DB	$\|A\|$	*ARE*
$P(State, Age, Sex)$	416	0.2111
$P(State, Labor\_status, Sex)$	208	0.1470
$P(Age, Labor\_status, Sex)$	16	0.1439
$\widehat{P}(State, Age, Labor\_status, Sex)$	832	0.1186

Table 7
ARE of $\widehat{I}(State, Labor\_status, Age, Sex)$ by applying the primary database $\widehat{I}(State, Labor\_status, Age)$ to proxy databases.

Proxy DB	$\|A\|$	*ARE*
$P(State, Age, Sex)$	416	0.2389
$P(State, Labor\_status, Sex)$	208	0.1909
$P(Age, Labor\_status, Sex)$	16	0.1827
$\widehat{P}(State, Age, Labor\_status, Sex)$	832	0.1625

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

consider different data sets, where $DB_{PR1} = Income(Region, Education\_level)$ is the primary database, which represents the total income (for the sake of brevity, we use simply "income") of households by *Region* and *Education_level* and multiple proxy databases shown below that represent the number of households by *Region*, *Education_level*, *Race*, *Age*, and *Tenure*. The reason for using another example here is to show better the effect of applying PP on the IPFP procedure. In this example, there are five dimensions, three target dimensions, and two non-common dimensions, whereas in the previous example all dimensions were target dimensions. *Group 1*

- $DB_{PX1} = Household(Region, Age, Education\_level)$
- $DB_{PX2} = Household(Region, Race, Tenure)$
- $DB_{PX3} = Household(Education\_level, Tenure)$

Suppose that the target database is *Income* by *Region*, *Education_level*, and *Tenure*, where the cardinalities of the dimensions are:

$$|\mathsf{Region}| = 4, |\mathsf{Race}| = 4, |\mathsf{Age}| = 7, |\mathsf{Education\_level}| = 9, \ \textit{and} \ |\mathsf{Tenure}| = 3.$$

Note that *Race* and *Age* are non-common dimensions. According to the maximum entropy principles discussed in the previous sections, we have to estimate first $Household(Region, Race, Age, Education\_level, Tenure)$ and then apply linear indirect estimation method to this and the primary database to estimate $DB_T = Income(Region, Education\_level, Tenure)$. Once the estimate of $\widehat{H}ousehold(Region, Race, Age, Education\_level, Tenure)$ is obtained by IPFP, we can estimate the target database by the F method as follows:

$$\widehat{I}[\mathsf{F}](Region, Race, Age, Education\_level, Tenure) = Income(Region, Education\_level) \times \frac{\widehat{H}ousehold(Region, Race, Age, Education\_level, Tenure)}{\sum_{Race,Age,Tenure} \widehat{H}ousehold(Region, Race, Age, Education\_level, Tenure)}$$

and then aggregate non-common dimensions *Race*, *Age* to obtain the target database:

$$\widehat{I}ncome(Region, Education\_level, Tenure) = \sum_{Race,Age} \widehat{I}ncome[\mathsf{F}](Region, Race, Age, Education\_level, Tenure)$$

The average relative error of the estimates is 0.207456 and shown in Table 8 (see first row, fourth column). If we preaggregate first the non-common dimensions over $\widehat{H}ousehold(Region, Race, Age, Education\_level, Tenure)$ as follows:

$$\widehat{H}ousehold(Region, Education\_level, Tenure) = \sum_{Race,Age} \widehat{H}ousehold(Income(Region, Race, Age, Education\_level, Tenure)$$

and then apply the linear indirect estimation method to estimate the target database as follows:

$$\widehat{I}ncome[\mathsf{PP}](Region, Education\_level, Tenure) = Income(Region, Education\_level) \times \frac{\widehat{H}ousehold(Region, Education\_level, Tenure)}{\sum_{Tenure} \widehat{H}ousehold(Region, Education\_level, Tenure)}$$

the average relative error is the same as ARE by the F method.

Overall, once we estimate the maximum entropy of the proxies (i.e., $\widehat{H}(Region, Race, Age, Education\_level, Tenure)$) by the F method, we can estimate the target database by the PP method, according to step (7) of PROCEDURE. Note that the average relative error of the estimate of the target database by applying the proxies $Household(Region, Race, Tenure)$ and $Household(Education\_level, Tenure)$ to the primary database $Income(Region, Education\_level)$, without computing the maximum entropy (by IPFP), are 0.212084 and 0.218138, respectively. In fact, they are higher than the ARE of the estimate of the target database calculated by applying the estimate of the number of household by the IPFP as shown above.

Table 8
ARE and time of computations.

Group	*Method*	*ARE of* $\widehat{H}$	*ARE of* DB_T	*Time* (s)
(1)	*F*	0.013960	0.207456	6.900
	PP	0.014300	0.207466	1.542
(2)	*F*	0.055471	0.206925	1.020
	PP	0.055471	0.206925	0.441

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

We emphasize that the PP method is only used in the estimation of the target database. Now, the question is whether we can use this method to estimate the maximum entropy by the IPFP procedure. In other words, can we apply the PP method or pre-aggregate first the non-common dimensions over source databases and then apply the IPFP procedure? Do we achieve, in this way, the same results as applying first the F method in IPFP procedure and then aggregate the non-common dimensions? This question is addressed through the data set labeled by *Group 1* and shown above and the following *Group 2* of proxy databases.

Group 2

- $DB_{PX1} = Household(Region, Age)$
- $DB_{PX2} = Household(Region, Race, Tenure)$
- $DB_{PX3} = Household(Education\_level, Tenure)$

We focus first on *Group 1* of proxy databases and apply the F method for running the IPFP procedure. Accordingly, $\widehat{H}ousehold(Region, Race, Age, Education\_level, Tenure)$ is computed. The number of iteration cycles needed to achieve the convergence and the relative execution time are shown in Table 9.

Then, we apply the PP method. In this case, we pre-aggregate first the dimensions *Age* and *Race* in DB_{PX1} and DB_{PX2}, respectively, and then run the IPFP procedure. In Table 9, we note that the convergence is achieved in 3 iteration cycles, while in the case of the F method the convergence is obtained by 6 iteration cycles. Consequently, the execution time of IPFP by the F method is 6.319, while by the PP method is 1.420. Regarding the accuracy, we note that the average relative error of the estimate $\widehat{H}(Region, Education\_level, Tenure)$ are not equal. The case of the F method (see Table 8, third column) is better than the average relative error of the same estimate by applying the PP method (0.013960 vs. 0.014300). Accordingly, the estimate of the target database obtained by the application of the linear indirect estimation method on the primary database and the two estimates of $\widehat{H}(Region, Education\_level, Tenure)$ by the F and PP methods mentioned above are also different, i.e., 0.207456 vs. 0.207466, (see Table 8, fourth column).

Now, we perform the same experiments over the second group of proxies (*Group 2*) in order to estimate the target database $Income(Region, Education\_level, Tenure)$, where *Race* and *Age* are non-common dimensions. As we mentioned above, in order to apply the PP method, we need to pre-aggregate *Age* in DB_{PX1} and *Race* in DB_{PX2}, respectively, as shown below. Hence, $DB_{PX1} = Household(Region)$ is a marginal of DB_{PX2} and as anticipated in Section 3, it is redundant. Thus, IPFP is applied to DB_{PX2}, and DB_{PX3}, and it consists of one iteration.

- $DB_{PX1} = Household(Region)$
- $DB_{PX2} = Household(Region, Tenure)$
- $DB_{PX3} = Household(Education\_level, Tenure)$

We note that applying the IPFP procedure to estimate $\widehat{H}ousehold(Region, Education\_level, Tenure)$ by the F method and the PP method give the same results in this case, i.e., 0.055471 (see Table 8, third column), and the average relative error of the estimates is also the same, i.e., 0.206925. In Table 9, the number of iteration cycles to achieve the convergence and time of executions of IPFP are shown, as well.

What is the difference between these two cases? It turns out that the difference between the above mentioned groups is related to the schemes of databases. Specifically, the hypergraph formed by the databases over their dimensions in *Group 1* is cyclic, while in *Group 2* is acyclic. In order to identify cycles in the hypergraph, one can use the well-known Graham Reduction algorithm [10]. We note that the basic concepts of hypergraph theory as well as the definitions of an acyclic hypergraph to implement the IPFP procedure are discussed in [2], where the authors combine a tree-implementation of the IPFP with an application of the principle of the divide-and-conquer. Based on these studies, and some additional experiments performed over different data sets, we conjecture that only if the schemes of the databases are acyclic, then the PP method can be applied in IPFP. However, we have no proof for that. The proof of this conjecture is a future challenge. As a final remark, we emphasize that cyclic/acyclic condition of the schema of the source databases has no effect on the principles of the maximum entropy and in both cases the estimation of the maximum entropy is provided by IPFP. These conditions are considered in order to estimate efficiently the maximum entropy. Accordingly, the main contribution of applying the PP method in IPFP, when the schemes of source databases are acyclic, consists of saving a significant number of computations and running time of the procedure.

Table 9
Application of IPFP in groups.

Group	*Method*	$\|A\|$	*Iteration of cycles*	*Time* (s)
(1)	*F*	3024	6	6.319
	PP	108	3	1.420
(2)	*F*	3024	3	0.952
	PP	108	1	0.341

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

Table 10
ARE of $\widehat{Income}$ computed from aggregate primary databases.

Primary DB	*Proxy DB*	*Target DB*	*ARE*
Income(*State*, *Age*)	*Population*(*State*, *Age*, *Sex*)	$\widehat{Income}$(*State*, *Age*, *Sex*)	0.72914
Income(*State*)	*Population*(*State*, *Age*, *Sex*)	$\widehat{Income}$(*State*, *Age*, *Sex*)	1.57797
Income(*Division*, *Age*)	*Population*(*Division*, *Age*, *Sex*)	$\widehat{Income}$(*Division*, *Age*, *Sex*)	0.35277
Income(*Division*)	*Population*(*Division*, *Age*, *Sex*)	$\widehat{Income}$(*Division*, *Age*, *Sex*)	0.63323
Income(*Region*, *Age*)	*Population*(*Region*, *Age*, *Sex*)	$\widehat{Income}$(*Region*, *Age*, *Sex*)	0.17991
Income(*Region*)	*Population*(*Region*, *Age*, *Sex*)	$\widehat{Income}$(*Region*, *Age*, *Sex*)	0.49911

6. Roll-up and drill-down

In this section, we examine the results obtained in the previous sections for the cases of roll-up and drill-down operations. We refer to the notation used in [15], and recall them next. We use the notation A_P^t and A_Q^t to represent two different dimension-levels in the category hierarchy of the same dimension t of DB_P and DB_Q. For example, *State* → *Division* are two dimension-levels in the dimension *Geographical_area* and *Date* → *Month* are two dimension-levels in the dimension *Time*. We use the notation for lower and higher category levels as $A_P^{t,L} \rightarrow A_P^{t,H}$ of target dimension A_P^t. Similarly, $A_Q^{t,L} \rightarrow A_Q^{t,H}$ of target dimension A_Q^t. Finally, we use the notation $A_P^{T'} = A_P^{T^{C'}} \cup A_P^{T^{\overline{C'}}}$ to represent the remaining target dimensions not involved in the roll-up or drill-down operation. Thus, $A_P^T = A_P^{T'} \cup A_P^{t,L}$. Similarly, $A_Q^T = A_Q^{T'} \cup A_Q^{t,L}$, where $A_Q^{T'} = A_Q^{T^{C'}} \cup A_Q^{T^{\overline{C'}}}$. This notation is used in the following definitions and theorems.

6.1. Roll-up

We consider the accuracy of estimates when multiple primary databases are aggregated over a dimension of a given classification hierarchy. For instance, let *State* be a dimension level in the dimension *Geographical_area* defined by three levels: *State* → *Division* → *Region*. Note that, according to *US Census Bureau*,[3] United States territory is subdivided into 9 divisions (i.e., New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Atlantic, West South Atlantic, Mountain, and Pacific), and 4 regions (Northeast, Midwest, South, West).

Consider the primary databases shown in Table 10, first column, where for each pair of primary databases, the proxy database and the target database are shown. Each pair of primary databases are defined at a certain level of the dimension *Geographical_area*. Note that the estimate of the target database is more accurate when the primary database with the largest number of cells in common with target database and proxy database is selected. For example, the accuracy of the estimate by using *Income*(*State*, *Age*) is better than the accuracy of estimate by using *Income*(*State*). Similar results are obtained by the remaining pairs of primary databases, and they again confirm the results discussed in the previous sections and proved by theorems shown in Section 4.

We note also that the estimate by applying *Income*(*Region*, *Age*) is more accurate than the estimates obtained by *Income*(*Division*, *Age*) and *Income*(*State*, *Age*), i.e., $\widehat{Income}$(*Region*, *Age*, *Sex*) is more accurate than $\widehat{Income}$(*Division*, *Age*, *Sex*), and $\widehat{Income}$(*State*, *Age*, *Sex*). Similarly, the estimate by *Income*(*Region*) is more accurate than the estimates by *Income*(*Division*), and *Income*(*State*). Overall, the aggregation of small cells into larger cells avoid the distortion in the distribution of the measure by small cells. Therefore, the accuracy of estimates is higher. This is proved by the next theorem, where we show that the ARE of estimate by applying a given primary database defined over a set of dimensions is higher than the ARE of estimate by applying the same primary database aggregated over a given dimension by roll-up operation. In order to prove this theorem, we use the notations introduced in the following definition.

Definition 6.1. Let $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right), M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$ be primary summary databases, and let $M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right), M_{Q_H}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H}\right)$ be proxy databases, where $A_P^{t,L} \rightarrow A_P^{t,H}, A_Q^{t,L} \rightarrow A_Q^{t,H}$ and $A_P^{t,L} = A_Q^{t,L}, A_P^{t,H} = A_Q^{t,H}$. We define $\widehat{M}$ to be the estimation result of the target database over the primary database $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)$ using $M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$. Similarly, we define $\widehat{\widehat{M}}$ to be the estimation result of target database over the primary database $M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$ using $M_{Q_H}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H}\right)$. The precise expressions for $\widehat{M}$ and $\widehat{\widehat{M}}$ using the PP method are provided below. According to this method, the source databases are aggregated first over non-common dimensions as follows:

[3] http://www.census.gov/geo/www/us_regdiv.pdf.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,L}\right) = \sum_{A_P^{\overline{C}}} M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,L}\right)$$

$$M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,H}\right) = \sum_{A_P^{\overline{C}}} M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,H}\right)$$

$$M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right) = \sum_{A_Q^{\overline{C}}} M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right)$$

$$M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right) = \sum_{A_Q^{\overline{C}}} M_{Q_H}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)$$

then, the linear indirect estimation is applied:

$$\begin{aligned}
\widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right) &= \sum_{A_Q^C} \widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^C, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right) \\
&= \sum_{A_Q^C}\left(M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,L}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right)}{\sum_{A_Q^{T^{\overline{C}'}}, A_Q^{t,L}} M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right)}\right) \\
&= \sum_{A_Q^C}\left(M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,L}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right)}{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}\right)}\right)
\end{aligned} \tag{3}$$

$$\begin{aligned}
\widehat{\widehat{M}}_{P_H}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right) &= \sum_{A_Q^C} \widehat{\widehat{M}}_{P_H}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^C, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right) \\
&= \sum_{A_Q^C}\left(M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,H}\right)\frac{M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)}{\sum_{A_Q^{T^{\overline{C}'}}, A_Q^{t,H}} M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)}\right) \\
&= \sum_{A_Q^C}\left(M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,H}\right)\frac{M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)}{M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}\right)}\right)
\end{aligned} \tag{4}$$

Theorem 6.1. *Let* $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,L}\right), M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_P^{t,H}\right)$ *be primary databases, and let* $M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right), M_{Q_H}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)$ *be proxy databases, where* $A_P^{t,L} \to A_P^{t,H}$, $A_Q^{t,L} \to A_Q^{t,H}$ *and* $A_P^{t,L} = A_Q^{t,L}$, $A_P^{t,H} = A_Q^{t,H}$. *The estimator of target database* $\widehat{\widehat{M}}_{P_H}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,H}\right)$ *gives a more accurate result than the estimator* $\widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C}'}}, A_Q^{T^{\overline{C}'}}, A_Q^{t,L}\right)$.

Proof. We show $ARE_{\widehat{\widehat{M}}_{P_H}} < ARE_{\widehat{M}_{P_L}}$ in Appendix C.1. □

Now consider estimating the target database *Income*(*Region*, *Age*, *Sex*) from the primary database *Income*(*State*, *Age*) and the proxy database *Population*(*State*, *Age*, *Sex*). The estimate can be computed in two ways. One way is to apply first the PP method to estimate $\widehat{I}ncome(State, Age, Sex)$ and then perform roll-up on the *State* dimension to achieve $\widehat{I}ncome(Region, Age, Sex)$. The other way is to apply first roll-up operation on the *State* dimension in the primary database

Table 11
ARE of $\widehat{I}ncome(Region, Age, Sex)$ computed by roll-up operation.

Primary databases	*ARE*(*Roll-up first then PP*)	*ARE* (*PP first then Roll-up*)
Income(*State*, *Age*)	0.17991	0.17690
Income(*State*, *Sex*)	0.49082	0.52792
Income(*State*)	0.49911	0.53672

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

and the proxy database (i.e., from $Income(State, Age), Population(State, Age, Sex)$ we obtain $Income(Region, Age)$, $Population(Region, Age, Sex)$) and then compute the estimate by the PP method. Obviously the second solution saves a number of computations but the question is which one achieves better accuracy for the estimate $\widehat{I}ncome(Region, Age, Sex)$.

In Table 11, the accuracy results shown in the first row show that the first way is better than the second one. Repeating this procedure on $Income(State, Sex)$ and $Income(Sex)$ to estimate $\widehat{I}ncome(Region, Age, Sex)$, the results by applying first the roll-up and then the PP method, i.e., the second way, are more accurate with respect to the estimate by the first way. It turns out that in similar cases, we cannot choose a priori which one of the two solutions (applying first roll-up then the PP method or viceversa) can achieve the better results. We conjecture that this probably depends on the distributions of measure values and on the dependency of the measure on dimensions, for which we have no proof.

6.2. Drill-down

The disaggregation over the category hierarchy, that is referred to as drill-down, occurs when different categories of the same hierarchy appear in the dimensions of the source summary databases. For instance, consider the source databases $Income(Region, Age)$ and $Population(State, Age, Sex)$. The target database is $Income(State, Age, Sex)$. We need to drill-down the income from the *Region* to the *State* level by using the *Population* database as a proxy. It is generated as follows:

$$\widehat{I}ncome(State, Age, Sex) = Income(Region, Age)\frac{Population(State, Age, Sex)}{\sum_{Sex} Population(Region, Age, Sex)}$$

Note that the term in the denominator $Population(Region, Age, Sex)$ in the above expression is obtained by a roll-up operation on $Population(State, Age, Sex)$.

Now, suppose we have multiple primary databases as follows: $Income(Division, Sex)$, $Income(Region, Sex)$ and one proxy database $Population(State, Age, Sex)$. The target database is $Income(State, Age, Sex)$. The problem is which primary database do we choose to obtain more accurate results. The accuracy results of estimate are shown in Table 12, second column. We note that ARE by applying more aggregated primary database is higher, and therefore the estimates are less accurate. Specifically, ARE of the estimate obtained by applying $Income(Region, Sex)$ is higher than ARE of the estimate obtained by $Income(Division, Sex)$, i.e., 3.15404 vs. 2.92769. Similar results are obtained by applying primary databases $Income(Division)$ and $Income(Region)$ (see Table 12, fourth column). Obviously, given the proxy database $Population(State, Age, Sex)$, the accuracy of $\widehat{I}ncome(State, Age, Sex)$ computed by applying $Income(State, Sex)$ is higher than the estimates obtained by applying drill-down operation on primary databases $Income(Division, Sex)$, and $Income(Region, Sex)$ (see Table 12, first row). This is proved by the next theorem, where we show that the accuracy of the estimate by applying a given primary database defined over a set of dimensions is higher than the accuracy of estimate by applying the same primary database disaggregated over a given dimension and defined by drill-down. We emphasize that the drill-down operation can only be performed when the dimensions in the two source databases that are involved in the drill-down operation must belong to the same category hierarchy. Furthermore, the lower category must belong to the proxy database. That is, $A_Q^{t,L} \rightarrow A_P^{t,H}$. In order to prove this theorem, we use the notations introduced in the following definition.

Definition 6.2. Let $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)$, and $M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$ be primary summary databases, and let $M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$ be proxy database, where $A_Q^{t,L} \rightarrow A_P^{t,H}$ and $A_P^{t,L} = A_Q^{t,L}$. We define $\widehat{M}$ to be the estimation result of the target database over the primary database $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)$. Similarly, we define $\widehat{\widehat{M}}$ to be the estimation results of the target database over the primary database $M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$. The precise expressions for $\widehat{M}$ and $\widehat{\widehat{M}}$ using the PP method are provided below. First, the source databases are aggregated over non-common dimensions as follows:

$$M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right) = \sum_{A_P^{\overline{C}}} M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)$$

$$M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right) = \sum_{A_P^{\overline{C}}} M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$$

$$M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right) = \sum_{A_Q^{\overline{C}}} M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$$

Table 12
ARE of $\widehat{I}ncome(State, Age, Sex)$ by applying primary databases and proxy database $Population(State, Age, Sex)$.

Primary DB	ARE	Primary DB	ARE
$Income(State, Sex)$	1.08540	$Income(State)$	1.57797
$Income(Division, Sex)$	2.92769	$Income(Division)$	3.00845
$Income(Region, Sex)$	3.15404	$Income(Region)$	3.20183

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

then, the linear indirect estimation is applied:

$$\widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right) = \sum_{A_Q^C} \widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{C}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$$

$$= \sum_{A_Q^C}\left(M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)}{\sum_{A_Q^{T^{\overline{C'}}}, A_Q^{t,L}} M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)}\right) \tag{5}$$

$$= \sum_{A_Q^C}\left(M_{P_L}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)}{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}\right)}\right)$$

$$\widehat{\widehat{M}}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right) = \sum_{A_Q^C} \widehat{\widehat{M}}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{C}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$$

$$= \sum_{A_Q^C}\left(M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)}{\sum_{A_Q^{T^{\overline{C'}}}} M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H}\right)}\right) \tag{6}$$

$$= \sum_{A_Q^C}\left(M_{P_H}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)\frac{M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)}{M_{Q_H}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{t,H}\right)}\right)$$

Note that the term in the denominator is obtained using roll-up[4] *as follows:*

$$M_{Q_H}(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H}) = \mathscr{R}_{A_Q^{t,L} \to A_P^{t,H}}\left(M_{Q_L}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)\right)$$

Theorem 6.2. Let $M_{P_L}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L}\right)$, and $M_{P_H}\left(A_P^C, A_P^{\overline{C}}, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H}\right)$ be primary databases, and let $M_{Q_L}\left(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$ be proxy database, where $A_Q^{t,L} \to A_P^{t,H}$ and $A_Q^{t,L} = A_P^{t,L}$. The estimator of target database $\widehat{M}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$ gives more accurate result than the estimator $\widehat{\widehat{M}}_{P_L}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L}\right)$.

Proof. We show $ARE_{\widehat{M}_{P_L}} < ARE_{\widehat{\widehat{M}}_{P_L}}$ in Appendix C.2. □

7. Conclusions

A common technique of constructing a target database from summary databases in the case that such a result cannot be obtained from a single summary database, is to select a summary primary database that has the desired target measure and use a proxy database with a different measure to estimate the result. In this paper, we considered the following problem. Given multiple primary and multiple proxy summary databases (i.e., summarized from a large base data cube), we investigate the problem of selecting the databases that provide the most precise estimate for a target database. We prove that the primary and proxy databases with the largest number of cells in common with each other and with the target database provide more accurate results. Our methodology is based on the principles of information entropy. Based on these results, we proposed an algorithmic approach for determining the steps to select or compute the source databases from multiple summary databases. To describe the proposed algorithm and verify the theoretical results, several example databases were used, and experimental results derived. Finally, the accuracy results in cases where dimensions of source databases are defined over a hierarchical structure and roll-up and drill-down operations are needed to achieve the desired target results are investigated.

Appendix A. The linear indirect estimation

The main idea of this method stems from its use in geographical regions. According to this method, data from surveys of variables of interest at the national or regional level is used to obtain estimates at more geographically disaggregated levels

[4] The roll-up operator is denoted by $\mathscr{R}_{A_1 \to A_2}(M(A_1))$, where A_1 and A_2 represent two category levels of a category hierarchy. It applies the aggregation function COUNT or SUM to the measure $M(A_1)$, and gives as result $M(A_2)$ [15].

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

such as counties or other small areas. An indirect estimation calculates values of the variable of interest using available auxiliary (called *predictor* or *proxy*) data at the local level that are correlated with the variable of interest [6]. Formally, let i denote a small area. A target measure $Y(d)$ is provided over a set of dimensions d. $Y(d)$ was generated from $Y(d) = \sum_i Y(i,d)$. $Y(i,d)$ is no longer available. However, auxiliary information in the form of $X(i,d)$ is available. A linear indirect estimation of Y for small area i is defined by:

$$\widehat{Y}(i) = \sum_d \widehat{Y}(i,d) = \sum_d Y(d)\frac{X(i,d)}{X(d)}$$

where $X(d) = \sum_i X(i,d)$. $X(i,d)/X(d)$ represents the proportion of the population of small area i relative to the total population over set of dimensions d, and $\sum_i \widehat{Y}(i)$ must be equal to $\sum_d Y(d)$ [6].

Appendix B. Average relative error

A method that is commonly-used for measuring accuracy is the average relative error (*ARE*) [6]. Formally, the average relative error (*ARE*) is:

$$ARE = \frac{1}{m}\sum_{i=1}^{m}\frac{|\hat{v}_i - v_i|}{v_i}.$$

where $\hat{v}_i$ and v_i are, respectively, the estimated and precise (or base data) values, and m is the number of small areas for which estimated values were calculated.

Appendix C. Proofs for Section 6

In this section we prove Theorems 6.1 and 6.2. The first theorem shows that the estimator $\widehat{\widehat{M}}$ obtained by applying the roll-up operation over a given dimension in a set of dimensions is more accurate than the estimator $\widehat{M}$ defined over the same set of dimensions. The second theorem shows that the estimator $\widehat{M}$ defined over a set of dimensions is more accurate that the estimator $\widehat{\widehat{M}}$ obtained by applying the drill-down operation over a given dimension in the same set of dimensions.

C.1. Proof of Theorem 6.1

We show $ARE_{\widehat{\widehat{M}}_{P_{H}}} < ARE_{\widehat{M}_{P_L}}$.

$$\frac{1}{m}\sum_{j=1}^{m}\left|\frac{\widehat{\widehat{M}}_{P_{H_j}} - M_{P_{H_j}}}{M_{P_{H_j}}}\right| < \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{M}_{P_{L_i}} - M_{P_{L_i}}}{M_{P_{L_i}}}\right| \tag{7}$$

The equation above can be written by using Eqs. (3) and (4) as follows:

$$\begin{aligned}&\frac{1}{m}\sum_{j=1}^{m}\left|\sum_{A_Q^C}\left(\frac{M_{P_{H_a}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,H_\gamma}\right)M_{Q_{H_d}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,H_\gamma}\right)}{M_{Q_{H_d}}\left(A_Q^C, A_Q^{T^{C'}}\right)M_{P_{H_j}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,H_\gamma}\right)}\right) - 1\right| \\ &< \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{A_Q^C}\left(\frac{M_{P_{L_b}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_\tau}\right)M_{Q_{L_f}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_\tau}\right)}{M_{Q_{L_f}}\left(A_Q^C, A_Q^{T^{C'}}\right)M_{P_{L_i}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_\tau}\right)}\right) - 1\right|\end{aligned} \tag{8}$$

where $1 \leqslant a \leqslant |A_{P_H}|, 1 \leqslant b \leqslant |A_{P_L}|, 1 \leqslant d \leqslant |A_{Q_H}|, 1 \leqslant f \leqslant |A_{Q_L}|$ with $a,b,d,f < m$ and $a,b,d,f < n$, and $1 \leqslant \gamma \leqslant \mathscr{K}, 1 \leqslant \tau \leqslant \mathscr{T}$ with $\mathscr{K} < \mathscr{T}$. Let denote $|A_{P_H}| = \mathscr{A}, |A_{P_L}| = \mathscr{B}, |A_{Q_H}| = \mathscr{D}, |A_{Q_L}| = \mathscr{F}$.

According to the summarizability condition discussed in [9],[5] the following partitions are defined:

[5] This condition states that it is possible to obtain from the summary database defined at category level A_1 of a given hierarchy, another summary database defined at the higher level A_2 of the same hierarchy by using the roll-up function.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$
\begin{aligned}
&M_{P_{H_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H_1}\right) = M_{P_{L_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_s}\right)\\
&M_{P_{H_2}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H_2}\right) = M_{P_{L_{s+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_1}\right) + \cdots + M_{P_{L_r}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_r}\right)\\
&\ldots\\
&M_{P_{H_{\mathscr{A}}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H_{\mathscr{K}}}\right) = M_{P_{L_{z+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathscr{B}}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_{\mathscr{F}}}\right)
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&M_{P_{H_1}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_1}\right) = M_{P_{L_1}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_s}\right)\\
&M_{P_{H_2}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_2}\right) = M_{P_{L_{s+1}}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{P_{L_r}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_r}\right)\\
&\ldots\\
&M_{P_{H_{\mathscr{A}}}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_{\mathscr{K}}}\right) = M_{P_{L_{z+1}}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathscr{B}}}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{\mathscr{F}}}\right)
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&M_{Q_{H_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_1}\right) = M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_s}\right)\\
&M_{Q_{H_2}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_2}\right) = M_{Q_{L_{s+1}}}\left(A_Q^C, A_P^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{s+1}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_r}\right)\\
&\ldots\\
&M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_{\mathscr{K}}}\right) = M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{\mathscr{F}}}\right)
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&M_{Q_{H_1}}\left(A_Q^C, A_Q^{T^{C'}}\right) = M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}\right)\\
&M_{Q_{H_2}}\left(A_Q^C, A_Q^{T^{C'}}\right) = M_{Q_{L_{s+1}}}\left(A_Q^C, A_P^{T^{C'}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}\right)\\
&\ldots\\
&M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C, A_Q^{T^{C'}}\right) = M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \ldots + M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C, A_Q^{T^{C'}}\right)
\end{aligned}
\tag{12}
$$

According to Eqs. (10)–(12) the left hand side of Eq. (8) is defined as follows:

$$
\begin{aligned}
&\frac{1}{m}\left(\left|\sum_{A_Q^C}\left(\frac{M_{P_{H_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H_1}\right)M_{Q_{H_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_1}\right)}{M_{Q_{H_1}}\left(A_Q^C, A_Q^{T^{C'}}\right)M_{P_{H_1}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_1}\right)}\right) - 1\right| + \cdots +\right.\\
&\left.\left|\sum_{A_Q^C}\left(\frac{M_{P_{H_{\mathscr{A}}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,H_{\mathscr{K}}}\right)M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_{\mathscr{K}}}\right)}{M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C, A_Q^{T^{C'}}\right)M_{P_{Hm}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,H_{\mathscr{K}}}\right)}\right) - 1\right|\right)\\
&= \frac{1}{m}\left(\left|\sum_{A_Q^C}\left(\frac{\left(M_{P_{L_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_s}\right)\right)}{\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)}\right.\right.\right.\\
&\left.\left.\left.\frac{\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_s}\right)\right)}{\left(M_{P_{L_1}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_s}\right)\right)}\right) - 1\right|\right.\\
&+ \left|\sum_{A_Q^C}\left(\frac{\left(M_{P_{L_{s+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_{s+1}}\right) + \cdots + M_{P_{L_r}}\left(A_P^C, A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_P^{t,L_r}\right)\right)}{\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)}\right.\right.\\
&\left.\left.\frac{\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{s+1}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_r}\right)\right)}{\left(M_{P_{L_{s+1}}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_{s+1}}\right) + \cdots + M_{P_{L_r}}\left(A_P^{T^{C'}}, A_P^{T^{\overline{C'}}}, A_Q^{T^{\overline{C'}}}, A_Q^{t,L_r}\right)\right)}\right) - 1\right| +
\end{aligned}
$$

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$\cdots + \left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_{z+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathcal{J}}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{\mathcal{F}}}\right)\right)}{\left(M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_{\mathcal{F}}}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)} \right.\right.$$

$$\left.\left. \frac{\left(M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{Q_{L_{\mathcal{F}}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{\mathcal{F}}}\right)\right)}{\left(M_{P_{L_{z+1}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathcal{J}}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{\mathcal{F}}}\right)\right)} - 1 \right| \right)$$

The expression above is used to rewrite Eq. (8) as follows, which provides the proof of theorem:

$$\frac{1}{m}\left(\left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_s}\right)\right)}{\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)} \right.\right.\right.$$

$$\left.\left.\left. \frac{\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_s}\right)\right)}{\left(M_{P_{L_1}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_1}\right) + \cdots + M_{P_{L_s}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_s}\right)\right)} - 1 \right| \right.\right.$$

$$\left.\left. + \left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_{s+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{s+1}}\right) + \cdots + M_{P_{L_r}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_r}\right)\right)}{\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)} \right.\right.\right.$$

$$\left.\left.\left. \frac{\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{s+1}}\right) + \cdots + M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_r}\right)\right)}{\left(M_{P_{L_{s+1}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{s+1}}\right) + \cdots + M_{P_{L_r}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_r}\right)\right)} - 1 \right| + \right.\right.$$

$$\cdots$$

$$\left.\left. \left| \frac{\left(\sum_{A_Q^C} \left(M_{P_{L_{z+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathcal{J}}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{\mathcal{F}}}\right)\right)\right)}{\left(M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right) + \cdots + M_{Q_{L_{\mathcal{F}}}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)} \right.\right.\right.$$

$$\left.\left.\left. \frac{\left(M_{Q_{L_{z+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{Q_{L_{\mathcal{F}}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{\mathcal{F}}}\right)\right)}{\left(M_{P_{L_{z+1}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{z+1}}\right) + \cdots + M_{P_{L_{\mathcal{J}}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{\mathcal{F}}}\right)\right)} - 1 \right| \right)$$

$$< \frac{1}{n}\left(\left(\left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_1}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_1}\right)\right)\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_1}\right)\right)}{\left(M_{Q_{L_1}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)\left(M_{P_{L_1}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_1}\right)\right)} \right) - 1 \right| \right.\right.$$

$$\left.\left. + \cdots + \left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_s}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_s}\right)\right)\left(M_{Q_{L_s}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_s}\right)\right)}{\left(M_{Q_{L_s}}\left(A_P^C, A_P^{T^{C'}}\right)\right)\left(M_{P_{L_s}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_s}\right)\right)} \right) - 1 \right| \right) \right.$$

$$\left. + \left(\left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_{s+1}}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_{s+1}}\right)\right)\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{s+1}}\right)\right)}{\left(M_{Q_{L_{s+1}}}\left(A_Q^C, A_Q^{T^{C'}}\right)\right)\left(M_{P_{L_{s+1}}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_{s+1}}\right)\right)} \right) - 1 \right| \right.\right.$$

$$\left.\left. + \cdots + \left| \sum_{A_Q^C} \left(\frac{\left(M_{P_{L_r}}\left(A_P^C, A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_P^{t,L_r}\right)\right)\left(M_{Q_{L_r}}\left(A_Q^C, A_Q^{T^{C'}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_r}\right)\right)}{\left(M_{Q_{L_r}}\left(A_P^C, A_P^{T^{C'}}\right)\right)\left(M_{P_{L_r}}\left(A_P^{T^{C'}}, A_P^{\overline{T^{C'}}}, A_Q^{\overline{T^{C'}}}, A_Q^{t,L_r}\right)\right)} \right) - 1 \right| \right) \right)$$

$$+\cdots+\left(\left|\sum_{A_Q^C}\left(\frac{\left(M_{P_{L_{z+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,L_{z+1}}\right)\right)\left(M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{z+1}}\right)\right)}{\left(M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}}\right)\right)\left(M_{P_{L_{z+1}}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{z+1}}\right)\right)}\right)-1\right|\right.$$

$$\left.+\cdots+\left|\sum_{A_Q^C}\left(\frac{\left(M_{P_{L_{\mathscr{J}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,L_{\mathscr{J}}}\right)\right)\left(M_{Q_{L_{\mathscr{J}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{\mathscr{J}}}\right)\right)}{\left(M_{Q_{L_{\mathscr{J}}}}\left(A_P^C,A_P^{T^{C'}}\right)\right)\left(M_{P_{L_{\mathscr{J}}}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{\mathscr{J}}}\right)\right)}\right)-1\right|\right)\right)$$

C.2. Proof of Theorem 6.2

We show $ARE_{\widehat{M}_{P_L}} < ARE_{\widehat{\widehat{M}}_{P_L}}$.

$$\frac{1}{m}\sum_{j=1}^{m}\left|\frac{\widehat{M}_{P_{L_j}}-M_{P_{L_j}}}{M_{P_{L_j}}}\right| < \frac{1}{m}\sum_{j=1}^{m}\left|\frac{\widehat{\widehat{M}}_{P_{L_j}}-M_{P_{L_j}}}{M_{P_{L_j}}}\right|$$

The expression above can be written by using Eqs. (5) and (6) as follows:

$$\frac{1}{m}\sum_{j=1}^{m}\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_b}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,L_\tau}\right)M_{Q_{L_f}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_\tau}\right)}{M_{Q_{L_f}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_j}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_\tau}\right)}\right)\right)-1\right|$$

$$<\frac{1}{m}\sum_{j=1}^{m}\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_a}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,H_\gamma}\right)M_{Q_{L_f}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_\tau}\right)}{M_{Q_{H_d}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_\gamma}\right)M_{P_{L_j}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_\tau}\right)}\right)\right)-1\right| \tag{13}$$

Using the partition indicated in Eq. (9), and the following:

$$\begin{aligned}
&M_{Q_{H_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_1}\right)=M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_1}\right)+\cdots+M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_s}\right)\\
&M_{Q_{H_2}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_2}\right)=M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{s+1}}\right)+\cdots+M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_r}\right)\\
&\cdots\\
&M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_{\mathscr{J}}}\right)=M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{z+1}}\right)+\cdots+M_{Q_{L_{\mathscr{J}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{\mathscr{J}}}\right)
\end{aligned} \tag{14}$$

the formula at the right hand side of Eq. (13) is defined as follows:

$$\frac{1}{m}\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_1}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,H_1}\right)M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_1}\right)}{M_{Q_{H_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_1}\right)M_{P_{L_1}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_1}\right)}\right)\right)-1\right|\right.$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_1}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,H_1}\right)M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_s}\right)}{M_{Q_{H_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_1}\right)M_{P_{L_s}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_s}\right)}\right)\right)-1\right|\right)$$

$$+\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_2}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,H_2}\right)M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{s+1}}\right)}{M_{Q_{H_2}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_2}\right)M_{P_{L_{s+1}}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_{s+1}}\right)}\right)\right)-1\right|\right.$$

$$\left.+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_2}}\left(A_P^C,A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_P^{t,H_2}\right)M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_r}\right)}{M_{Q_{H_2}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_2}\right)M_{P_{L_r}}\left(A_P^{T^{C'}},A_P^{\overline{T^{C'}}},A_Q^{\overline{T^{C'}}},A_Q^{t,L_r}\right)}\right)\right)-1\right|\right)$$

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$+\cdots+\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_{\mathscr{A}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,H_{\mathscr{K}}}\right)M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}{M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_{\mathscr{K}}}\right)M_{P_{L_{z+1}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}\right)\right)-1\right|\right.$$

$$\left.+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{H_{\mathscr{A}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,H_{\mathscr{K}}}\right)M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}{M_{Q_{H_{\mathscr{J}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,H_{\mathscr{K}}}\right)M_{P_{L_{\mathscr{B}}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}\right)\right)-1\right|\right)\right)$$

Then, Eq. (13) can be written as follows, which provides the proof of theorem:

$$\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_1}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_1}\right)M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_1}\right)}{M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_1}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_1}\right)}\right)\right)-1\right|\right.$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_s}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_s}\right)M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}}\right)}{M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_s}\right)M_{P_{L_s}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_s}\right)}\right)\right)-1\right|\right)$$

$$+\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_{s+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{s+1}}\right)M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{s+1}}\right)}{M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_{s+1}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{s+1}}\right)}\right)\right)-1\right|\right.$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_r}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_r}\right)M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_r}\right)}{M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_r}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_r}\right)}\right)\right)-1\right|\right)$$

$$+\cdots+\left(\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_{z+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{z+1}}\right)M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}{M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_{z+1}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}\right)\right)-1\right|\right.$$

$$\left.+\cdots+\left|\left(\sum_{A_Q^C}\left(\frac{M_{P_{L_{\mathscr{B}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{\mathscr{F}}}\right)M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}{M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}}\right)M_{P_{L_{\mathscr{B}}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}\right)\right)-1\right|\right)$$

$$<\left(\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_1}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_1}\right)+\cdots+M_{P_{L_s}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_s}\right)\right)\right.\right.\right.$$

$$\left.\left.\left.\frac{M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_1}\right)}{\left(M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_1}\right)+\cdots+M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_s}\right)\right)M_{P_{L_1}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_1}\right)}\right)\right)-1\right|\right.$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_1}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_1}\right)+\cdots+M_{P_{L_s}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_s}\right)\right)\right.\right.\right.$$

$$\left.\left.\left.\frac{M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_s}\right)}{\left(M_{Q_{L_1}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_1}\right)+\cdots+M_{Q_{L_s}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_s}\right)\right)M_{P_{L_s}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_s}\right)}\right)\right)-1\right|\right)$$

$$+\left(\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_{s+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{s+1}}\right)+\cdots+M_{P_{L_r}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_r}\right)\right)\right.\right.\right.$$

$$\left.\left.\left.\frac{M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{s+1}}\right)}{\left(M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{s+1}}\right)+\cdots+M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_r}\right)\right)M_{P_{L_{s+1}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{s+1}}\right)}\right)\right)-1\right|$$

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_{s+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{s+1}}\right)+\cdots+M_{P_{L_r}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_r}\right)\right)\right.\right.$$
$$\left.\left.\frac{M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_r}\right)}{\left(M_{Q_{L_{s+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_1}\right)+\cdots+M_{Q_{L_r}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_r}\right)\right)M_{P_{L_r}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_r}\right)}\right)\right)-1\right|$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_{z+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{z+1}}\right)+\cdots+M_{P_{L_{\mathscr{F}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{\mathscr{F}}}\right)\right)\right.\right.$$
$$\left.\left.\frac{M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}{\left(M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{z+1}}\right)+\cdots+M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{\mathscr{F}}}\right)\right)M_{P_{L_{z+1}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{z+1}}\right)}\right)\right)-1\right|$$

$$+\cdots+\left|\left(\sum_{A_Q^C}\left(\left(M_{P_{L_{z+1}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{z+1}}\right)+\cdots+M_{P_{L_{\mathscr{F}}}}\left(A_P^C,A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_P^{t,L_{\mathscr{F}}}\right)\right)\right.\right.$$
$$\left.\left.\frac{M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}{\left(M_{Q_{L_{z+1}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{z+1}}\right)+\cdots+M_{Q_{L_{\mathscr{F}}}}\left(A_Q^C,A_Q^{T^{C'}},A_Q^{t,L_{\mathscr{F}}}\right)\right)M_{P_{L_{\mathscr{F}}}}\left(A_P^{T^{C'}},A_P^{T^{\overline{C'}}},A_Q^{T^{\overline{C'}}},A_Q^{t,L_{\mathscr{F}}}\right)}\right)\right)-1\right|$$

References

[1] P.B. Gibbons, Y. Matias, New sampling-based summary statistics for improving approximate query answers, ACM SIGMOD Int. Conference on Management of Data (1998) 331–342.
[2] J.-H. Badsberg, F.M. Malvestuto, An implementation of the iterative proportional fitting procedure by propagation trees, Computational Statistics and Data Analysis 37 (2001) 297–322.
[3] P.S. Bradley, U.M. Fayyad, J. Shanmugasundaram, Compressed data cubes for OLAP aggregate query approximation on continuous dimensions, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, United States, 1999, pp. 223–232.
[4] W.E. Deming, F.F. Stephan, On a least square adjustment of a sampled frequency table when the expected marginal totals are known, Annals of Mathematical Statistics 11 (1940) 427–444.
[5] J.M. Hellerstein, P.J. Haas, H.J. Wang, Online aggregation, ACM SIGMOD International Conference on Management of Data (1997) 171–182.
[6] M. Ghosh, J.N.K. Rao, Small area estimation: an appraisal, Statistical Science 9 (1994) 55–93.
[7] E. Jaynes, Where do we stand on maximum entropy?, in: R. Levine, M. Tribes (Eds.), The Maximum Entropy Formalism, MIT Press, Cambridge, MA, 1979, pp. 15–118.
[8] S. Kullback, Where Do We Stand on Maximum Entropy? Information Theory and Statistics, John Wiley and Sons, Inc., London, 1959.
[9] H.-J. Lenz, A. Shoshani, Summarizability in OLAP and statistical databases, in: Y.E. Ioannidis, D.M. Hansen (Eds.), Proceedings of Ninth International Conference on Scientific and Statistical Data Management (SSDBM), IEEE Computer Society, Olympia, WA, USA, 1997, pp. 132–143.
[10] D. Maier, The theory of relational databases – Chapter 13, Computer Science Press, 1983.
[11] M.F. Malvestuto, A universal-scheme approach to statistical databases containing homogeneous summary tables, ACM Transactions on Database Systems (TODS) 18 (4) (1993) 678–708.
[12] M.F. Malvestuto, E. Pourabbas, Customized answers to summary queries via aggregate views, in: Proceedings of 16th International Conference on Scientific and Statistical Database Management (SSDBM), IEEE Computer Society, Santorini Island, Greece, June 21–23, 2004, pp. 193–202.
[13] M.F. Malvestuto, E. Pourabbas, Local computation of answers to table queries on summary databases, in: Proceedings of 17th International Conference on Scientific and Statistical Database Management (SSDBM), Santa Barbara, CA, USA, June 27–29, 2005, pp. 263–270.
[14] T. Palpanas, N. Koudas, A. Mendelson, Using datacube aggregates for approximate querying and deviation detection, IEEE Transactions on Knowledge and Data Engineering 17 (11) (2005) 1465–1477.
[15] E. Pourabbas, A. Shoshani, Efficient estimation of joint queries from multiple OLAP databases, ACM Transactions on Database Systems (TODS) 32 (1) (2007).
[16] E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, in: Proceedings of ACM 11th International Workshop on Data Warehousing and OLAP, DOLAP 2008, Napa Valley, CA, USA, October 30, 2008, pp. 33–40.

Elaheh Pourabbas is a research scientist at the Istituto di Analisi dei Sistemi ed Informatica (IASI) "Antonio Ruberti" of the Italian National Research Council (http://www.iasi.cnr.it). She received her MS degree in electrical engineering from the University of Rome "La Sapienza" in 1992, and her Ph.D. degree in bioengineering from the University of Bologna in 1997. In 2005 she was awarded a Fulbright Fellowship in support of research carried out at the University of California, Lawrence Berkeley National Laboratory, Berkeley, USA. She served as referee of several international journals and conferences. Her research interests include query processing, data warehousing and OLAP, semantic web, and similarity reasoning.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010

Arie Shoshani is a senior staff scientist at Lawrence Berkeley National Laboratory. He joined LBNL in 1976, and currently heads the Scientific Data Management Group. He received his Ph.D. from Princeton University in 1969. His current areas of work include data models, temporal data, statistical and scientific database management, storage resource management and reservation, and grid storage middleware. Arie is also the director of a Scientific Data Management Center, funded by the Department of Energy (http://sdmcenter.lbl.gov). He published over 85 papers in referred journals and conferences. He holds jointly with colleagues a patent on a compression method for bitmap indexing, as well as receiving the 2008 R&D100 award for a very fast indexing method based on this patent. He served as an associate editor for the ACM Transactions on Database Systems, and was elected as Vice-President of the VLDB Endowment Board, and is the steering committee chair of the scientific and Statistical Data Base Management conference series since its inception (http://ssdbm.org). His home page is http://www.lbl.gov/arie.

Please cite this article in press as: E. Pourabbas, A. Shoshani, Improving estimation accuracy of aggregate queries on data cubes, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.08.010