Improving Estimation Accuracy of Aggregate Queries on Data Cubes

Elaheh Pourabbas
Institute of Systems Analysis and Computer Science "Antonio Ruberti"
National Research Council-CNR
Viale Manzoni 30
Rome, Italy
elaheh.pourabbas@iasi.cnr.it

Arie Shoshani*
Lawrence Berkeley National Laboratory
Mailstop 50B-3238
1 Cyclotron Road Berkeley
CA 94720 USA
shoshani@lbl.gov

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 which uses 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, that share the same summary measure, 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 for determining the steps to select or compute the source databases from multiple summary databases, which makes use of the principles of information entropy. We show that the source databases with the largest number of cells in common provide 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.

Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications; H.2.8 [**Database Applications**]: [Statistical Databases]

General Terms

Management, Theory, Experimentation

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 a 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), and was the focus of recent research activity (e.g., (Palpanas, Koudas & Mendelson 2005), (Pourabbas & Shoshani 2007)).

In (Pourabbas & Shoshani 2007), 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,Sex,Education_level)*. It is possible to estimate the target database *Income by (State,Sex)* by using the first database as the "primary" 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 that previous paper 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

*This work was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

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 (Pourabbas & Shoshani 2007). 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.

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(Labor_status,Age)*
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: $|State| = 52$, $|Age|=4$, $|Labor\_status|=2$, and $|Sex|=2$. Note that the two categories of *Labor_status* are *In_ Labor_Force* and *Not_in_Labor_Force* according to U.S. Census Bureau. Let *Income(State,Labor_status,Age,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.

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}

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. The problem is which databases should we choose from a given set of primary and proxy databases that provide more accurate estimation results.

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 includes 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 more accurate results.

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.

The problem addressed in this paper consists of the general problem labeled by *i* shown in Table 2. In (Pourabbas & Shoshani 2007), we studied the case *iv*. In this paper, we examine the first general case. The problems *ii*, *iii*, and *iv* are special cases of the problem *i* as well.

Table 2: *Problems*

Cases	*Primary DB*	*Proxy DB*
i	M *choice*	N *choice*
ii	M *choice*	N *given*
ii	1 *given*	N *choice*
iv	1 *given*	N *given*

1.2 Related Work

There was a significant amount of work in the literature on approximate query processing. In (Malvestuto 1993), 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 (Malvestuto & Pourabbas 2004), and (Malvestuto & Pourabbas 2005), 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 (Hellerstein, Haas & Wang 1997) 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) (Gibbons & Matias 1998) system provides approximate answers using small precomputed synopses of the underlying base data. In (Palpanas et al. 2005), 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, which achieve more accurate results. We prove formally that the source databases with the largest number of cells in common provide more accurate estimated results. 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 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. Section 6 contains the conclusions.

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

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

where X represents the set of tuples. 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 $X = \{X_1, , X_n\}$, 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 $\hat{P}(X)$ be the maximum entropy approximation to $P(X)$. The *cross-entropy* (or *relative entropy* or *Kullback-Leibler distance*) between $\hat{P}(X)$ and $P(X)$ measures the similarity of two distribution and is defined as follows:

$$D(\hat{P}, P) = \sum_X \hat{P}(X) log \frac{\hat{P}(X)}{P(X)} \quad (2)$$

Minimizing $D(\hat{P}, P)$ is the same as maximizing the entropy of P. The technique used to compute the maximum entropy estimate is *Iterative Proportional Fitting Procedure*-IPFP (Deming & Stephan 1940), 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, $1 \leq i \leq n$, 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 (Pourabbas & Shoshani 2007), 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: $DB_P = M_P(\{A_P^i \quad 0 < i \leq m\})$, $DB_Q = M_Q(\{A_Q^j \quad 0 < j \leq n\})$, and $DB_T = M_T(\{A_T^k \quad 0 < k \leq t\})$, 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: $DB_P = M_P(A_P^C, A_P^{\overline{C}}, A_P^{T^C}, A_P^{T^{\overline{C}}})$, and $DB_Q = M_Q(A_Q^C, A_Q^{\overline{C}}, A_Q^{T^C}, A_Q^{T^{\overline{C}}})$, 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 $\{A_T^k \quad 0 < k \leq t\}$. 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 $A_{Population}^{T^C} = A_{Income}^{T^C} = \emptyset$, $A_{Population}^{T^{\overline{C}}} = \{State\}$, $A_{Income}^{T^{\overline{C}}} = \emptyset$ are the non-common target dimensions, $A_{Population}^{C} = A_{Income}^{C} = \{Age, Sex\}$ are the common dimensions between the source summary databases, and $A_{Population}^{\overline{C}} = \{Race\}$, and $A_{Income}^{\overline{C}} = \{Labor\_status\}$ are the non-common dimensions. If the summary query expressed over the source databases is *Income(State,Age)*, then $A_T = \{State, Age\}$ and accordingly, $A_{Population}^{T^C} = A_{Income}^{T^C} = \{Age\}$, $A_{Population}^{T^{\overline{C}}} = \{State\}$, $A_{Income}^{T^{\overline{C}}} = \emptyset$, and $A_{Population}^{C} = A_{Income}^{C} = \{Sex\}$.

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 multiple proxy databases is needed for the target database and is not provided from primary databases. We recall the results discussed in (Pourabbas & Shoshani 2007) 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 the full cross-product of all dimensions of the source databases first 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 3 (where we use the symbols "*I*" and "*P*" to indicate *Income* and *Population*, respectively.) By multiplying the cardinalities of the dimensions we obtain the number of cells for each choice. As can be seen in Table 3, DB_{PR4} shares 416 cells for dimensions in common with the target database $Income(State, Labor\_status, Age, Sex)$. 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 Subsection 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 (Jaynes 1979). 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 3, the largest entropy is achieved by DB_{PR4}. This primary database also satisfies the three constraints of consistency conditions listed in Subsection 2.1. Concerning the proxy databases (see Table 4), 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.

Table 3: Primary databases

Primary DB	$\|A\|$	*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(Labor_status,Age)*	8	3.49	0.13815
DB_{PR4}=*I(State,Labor_status,Age)*	416	7.10	0.01623

Table 4: 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

The relative entropy (or loss of information) of the estimates by applying each primary database to DB_{PX4} is shown in Table 3, fourth column. Applying 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. Suppose, in Table 3, 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 Subsection 2.1). Thus, $Income(State, Labor\_status, Age)$ needs to be estimated. For this reason, we have to consider all three primary databases by applying IPFP to estimate $\hat{I}$*ncome(State,Labor_status, Age)*. This estimates 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(Labor_status,Age)*. 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 (Malvestuto & Pourabbas 2005), 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","L",and "A" indicate "*State*", "*Labor_status*", and "*Age*", respectively.)

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

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 mentioned proxy is DB_{PX4}, where $Pop(S,A,L) = \sum_{Sex} Pop(S,A,L,Sex)$.

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,L,A)*, but this proxy is not provided. Our solution is to estimate $\hat{P}(S, L, A, Sex)$ 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 $\hat{P}(State, Labor\_status, Age, Sex)$ given below:

$$\hat{P}(State, Labor\_status, Age, Sex) = P(State, Age, Sex)\frac{Pop(State,Labor\_status,Sex)}{Pop(State,Sex)}$$

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 Subsection 2.2, we can formulate the following definition and theorems.

Definition 1. Let $M_{P_k}(A^C_{P_k}, A^{\overline{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})$, $M_{P_l}(A^C_{P_l}, A^{\overline{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})$ be primary summary databases, and let $M_Q(A^C_Q, A^{\overline{C}}_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)$ be a proxy database. We define $\hat{M}_{P_k}$ to be the estimation result of the target database over the primary summary database $M_{P_k}(A^C_{P_k}, A^{\overline{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})$. Similarly, we define $\hat{M}_{P_l}$ to be the estimation result of target database over the primary database $M_{P_l}(A^C_{P_l}, A^{\overline{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})$. 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:

$$M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}) = \sum_{A^{\overline{C}}_{P_k}} M_{P_k}(A^C_{P_k}, A^{\overline{C}}_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})$$
$$M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}) = \sum_{A^{\overline{C}}_{P_l}} M_{P_l}(A^C_{P_l}, A^{\overline{C}}_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})$$
$$M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q) = \sum_{A^{\overline{C}}_Q} M_Q(A^C_Q, A^{\overline{C}}_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)$$

then, linear indirect estimation method is applied:

$$\hat{M}_{P_k}(A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}, A^C_Q, A^{T^{\overline{C}}}_Q) = M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})\frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_Q, A^{T^C}_Q)}$$

$$\hat{M}_{P_l}(A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}, A^C_Q, A^{T^{\overline{C}}}_Q) = M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})\frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_Q, A^{T^C}_Q)}$$

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

THEOREM 1. *Let $M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})$, $M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})$ be primary databases, and let $M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)$ be*

proxy database, where $|A_{P_l}| < |A_{P_k}| < |A_T|$, $A^{\mathcal{C}}_{P_l} \subset A^{\mathcal{C}}_{P_k}$, *and* $\mathcal{C}$ *represents common and common-target dimension groups. Let* $\hat{M}_{P_k}$ *and* $\hat{M}_{P_l}$ *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 $\hat{M}_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}\prime}}_{P_k}, A^{T^{\overline{C}}}_{Q})$ and $\hat{M}_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}\prime}}_{P_l}, A^{T^{\overline{C}}}_{Q})$ with respect to $M_P(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}, A^{T^{\overline{C}}}_{Q})$ be defined according to expressions:

$$D(\hat{M}_{P_k}, M_P) = \sum \hat{M}_{P_k} log \frac{\hat{M}_{P_k}}{M_P}$$
$$= \sum \left(\left(M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}) \frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_{P_k}, A^{T^C}_{P_k})} \right) \right.$$
$$\left. log \frac{M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}) \frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_{P_k}, A^{T^C}_{P_k})}}{M_P(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}, A^{T^{\overline{C}}}_{Q})} \right)$$

$$D(\hat{M}_{P_l}, M_P) = \sum \hat{M}_{P_l} log \frac{\hat{M}_{P_l}}{M_P}$$
$$= \sum \left(\left(M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}) \frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_{P_l}, A^{T^C}_{P_l})} \right) \right.$$
$$\left. log \frac{M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}) \frac{M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q)}{M_Q(A^C_{P_l}, A^{T^C}_{P_l})}}{M_P(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}, A^{T^{\overline{C}}}_{Q})} \right)$$

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

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

Setting $\mathcal{F} = M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q) \frac{M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}) M_Q(A^C_{P_l}, A^{T^C}_{P_l})}{M_Q(A^C_{P_k}, A^{T^C}_{P_k}) M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})}$ and $\mathcal{G} = \frac{M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}) M_Q(A^C_{P_k}, A^{T^C}_{P_k})}{M_Q(A^C_{P_l}, A^{T^C}_{P_l}) M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})}$,

$$D(\hat{M}_{P_l}, M_P) - D(\hat{M}_{P_k}, M_P) =$$
$$\sum \mathcal{F}\mathcal{G} log \frac{M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}) M_Q(A^C_{P_k}, A^{T^C}_{P_k})}{M_Q(A^C_{P_l}, A^{T^C}_{P_l}) M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})} = \sum \mathcal{F}\mathcal{G} log \mathcal{G}$$

Since $\sum \mathcal{F}\mathcal{G} = \sum M_Q(A^C_Q, A^{T^C}_Q, A^{T^{\overline{C}}}_Q) = 1$, and according to Theorem 3.1 (The theorem states the relative entropy obtained from distributions of the observations is positive, see Chapter 2) in (Kullback 1959)

$$\sum \mathcal{G} log \mathcal{G} = D\left(M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l}), M_Q(A^C_{P_l}, A^{T^C}_{P_l}) \right) +$$
$$D\left(M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k}), M_Q(A^C_{P_k}, A^{T^C}_{P_k}) \right)$$

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

$$\frac{M_{P_l}(A^C_{P_l}, A^{T^C}_{P_l}, A^{T^{\overline{C}}}_{P_l})}{M_Q(A^C_{P_l}, A^{T^C}_{P_l})} = \frac{M_{P_k}(A^C_{P_k}, A^{T^C}_{P_k}, A^{T^{\overline{C}}}_{P_k})}{M_Q(A^C_{P_k}, A^{T^C}_{P_k})}$$

□

Definition 2. Let $M_P(A^C_P, A^{\overline{C}}_P, A^{T^C}_P, A^{T^{\overline{C}}}_P)$, be primary database, and let $M_{Q_k}(A^C_{Q_k}, A^{\overline{C}}_{Q_k}, A^{T^C}_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$, $M_{Q_l}(A^C_{Q_l}, A^{\overline{C}}_{Q_l}, A^{T^C}_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$ be proxy databases. We define $\hat{M}_{P_k}$ to be the estimation result of the target database by applying the primary database to $M_{Q_k}(A^C_{Q_k}, A^{\overline{C}}_{Q_k}, A^{T^C}_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$. Similarly, we define $\hat{M}_{P_l}$ to be the estimation result of target database by applying the primary database to $M_{Q_l}(A^C_{Q_l}, A^{\overline{C}}_{Q_l}, A^{T^C}_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$. The expressions of the estimators above are defined by applying the PP method as follows:

$$\hat{M}_{P_k}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$$
$$= M_P(A^C_P, A^{T^C}_P, A^{T^{\overline{C}}}_P) \frac{M_{Q_k}(A^C_{Q_k}, A^{T^C}_{Q_k}, A^{T^{\overline{C}}}_{Q_k})}{M_{Q_k}(A^C_{Q_k}, A^{T^C}_{Q_k})}$$
$$\hat{M}_{P_l}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$$
$$= M_P(A^C_P, A^{T^C}_P, A^{T^{\overline{C}}}_P) \frac{M_{Q_l}(A^C_{Q_l}, A^{T^C}_{Q_l}, A^{T^{\overline{C}}}_{Q_l})}{M_{Q_l}(A^C_{Q_l}, A^{T^C}_{Q_l})}$$

THEOREM 2. $M_P(A^C_P, A^{T^C}_P, A^{T^{\overline{C}}}_P)$*, be primary database, and let* $M_{Q_k}(A^C_{Q_k}, A^{T^C}_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$, $M_{Q_l}(A^C_{Q_l}, A^{T^C}_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$ *be proxy databases, where* $|A_{Q_l}| < |A_{Q_k}|$, $A^{\mathcal{C}}_{Q_l} \subset A^{\mathcal{C}}_{Q_k}$*. Let* $\hat{M}_{P_k}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$ *and* $\hat{M}_{P_l}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$ *be the estimate of the target database obtained by applying the primary database* $M_P(A^C_P, A^{T^C}_P, A^{T^{\overline{C}}}_P)$ *to* M_{Q_k} *and* M_{Q_l}*, respectively. The estimate* $\hat{M}_{P_k}$ *is more accurate than the estimate* $\hat{M}_{P_l}$*.*

Proof Let the relative entropy of $\hat{M}_{P_k}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_k}, A^{T^{\overline{C}}}_{Q_k})$ and $\hat{M}_{P_l}(A^{T^C}_P, A^{T^{\overline{C}}}_P, A^C_{Q_l}, A^{T^{\overline{C}}}_{Q_l})$ with respect to $M_P(A^C_P, A^{T^C}_P,$

$A_P^{T^{\overline{C}}}, A_Q^{T^{\overline{C}}}$) be defined according to the following expressions:

$$D(\hat{M}_{P_k}, M_P) = \sum \hat{M}_{P_k} log \frac{\hat{M}_{P_k}}{M_P}$$
$$= \sum \left(\left(M_P(A_P^C, A_P^{T^C}, A_P^{T^{\overline{C}}}) \frac{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\overline{C}}})}{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C})} \right) log \frac{M_P(A_P^C, A_P^{T^C}, A_P^{T^{\overline{C}}}) \frac{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\overline{C}}})}{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C})}}{M_P(A_P^C, A_P^{T^C}, A_P^{T^{\overline{C}}}, A_{Q_k}^{T^{\overline{C}}})} \right)$$

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

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

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

Setting

$\mathcal{F} = M_P(A_P^C, A_P^{T^C}, A_P^{T^{\overline{C}}}) \frac{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\overline{C}}}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C})}{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\overline{C}}})}$

and $\mathcal{G} = \frac{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\overline{C}}})}{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\overline{C}}}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C})}$,

$$D(\hat{M}_{P_l}, M_P) - D(\hat{M}_{P_k}, M_P) =$$
$$\sum \mathcal{F}\mathcal{G} log \frac{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C}, A_{Q_l}^{T^{\overline{C}}})}{M_{Q_k}(A_{Q_k}^C, A_{Q_k}^{T^C}, A_{Q_k}^{T^{\overline{C}}}) M_{Q_l}(A_{Q_l}^C, A_{Q_l}^{T^C})} = \sum \mathcal{F}\mathcal{G} log \mathcal{G}$$

Similar to Theorem 1, $\sum \mathcal{F}\mathcal{G} log \mathcal{G} > 0$ 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 by 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.

PROCEDURE
Input: Given target database DB_T, multiple primary databases DB_{PRi} with $1 \leq i \leq n$ and multiple proxy databases DB_{PXj} $1 \leq j \leq m$ 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 $A_{PX}^{T^{\overline{C}}}$;
(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 $A_{PX} = A_{PX}^{T^{\overline{C}}} \cup A_{PR}$;
(6) If there is no such proxy database, consider proxy databases that have $A_{PX} = A_{PX}^{T^{\overline{C}}}$, with additional dimensions such that:
(a) if non-common, pre-aggregate dimensions;
(b) if common, apply IPFP to multiple proxy databases;
PART 3- ESTIMATION OF THE TARGET DATABASE
(7) Apply linear indirect estimation method to DB_{PR}, and DB_{PX}.

5. EXPERIMENTAL RESULTS

We discuss the experimental results of the application of our algorithmic approach to the four cases introduced in Subsection 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 5, first 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 this 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, we note the accuracy results in *Case 2* is close to that of *Case 4*. Similarly, the accuracy of results in *Case 1* is close to that of *Case 3*. With respect to *Case 1*, the accuracy of *Case 3* is better that *Case 2*. It seems that the estimation of the proxy database does not effect significantly the accuracy of the results. But, this is not the case of the estimation of the primary database (see the accuracy of *Case 1* and *Case 2*). Obviously, the accuracy of *Case 4* is worse than the other cases.

In addition, we compare some accuracy results of the estimates. Specifically, in Table 6, we compare the accuracy results of the estimate of target database by applying each primary database to $P(State, Labor\_status, Age, Sex)$ and the estimate of the primary database computed according to step (3) of the proposed procedure. Table 7 illustrates

Table 5: Accuracy results of selected primary and proxy databases in four cases

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

Table 6: ARE of $\hat{I}(State, Labor\_status, Age, Sex)$ 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
$\hat{I}(State, Age, Labor\_status)$	416	0.1464

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 8 shows the accuracy results of the estimate of the target database by applying the estimated primary database $\hat{I}(State, Labor\_status, Age)$ to each given proxy database and the estimated proxy database $\hat{P}(State, Labor\_status, Age, Sex)$.

6. CONCLUSIONS

Given multiple primary and multiple proxy databases summarized over a large base cube database, we investigate the problem of selecting the source summary databases that provide the most precise estimate for a target database. The databases in each set share the same summary measure. We show that the primary and proxy databases with the largest number of cells in common 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 such proposed algorithm, some example databases were used, and experimental results for them have been demonstrated.

Table 7: ARE of $\hat{I}(State, Labor\_status, Age, Sex)$ 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
$\hat{P}(State, Age, Labor\_status, Sex)$	832	0.1186

Table 8: ARE of $\hat{I}(State, Labor\_status, Age, Sex)$ by applying the primary database $\hat{I}(State, Labor\_status, Age)$ 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
$\hat{P}(State, Age, Labor\_status, Sex)$	832	0.1625

7. REFERENCES

Deming, W. E. & Stephan, F. F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics* **11**: 427–444.

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: An appraisal, *Statistical Science* **9**: 55–93.

Gibbons, P. & Matias, Y. (1998). New sampling-based summary statitistics for improving approximate query answers, *Proceedings of ACM SIGMOD International Conference on Management of Data (ACM SIGMOD),* Seattle, Washington L. M. Haas and A. Tiwary, Eds., ACM Press, pp. 232–243.

Hellerstein, J., Haas, P. & Wang, H. (1997). Online aggregation, *Proceedings ACM SIGMOD International Conference on Management of Data (ACM SIGMOD),* Tucson, Arizona, USA, Joan Peckham Eds., ACM Press, pp. 171–182.

Jaynes, E. (1979). Where do we stand on maximum entropy?, *The Maximum Entropy Formalism,* R. Levine and M. Tribes Eds., MIT Press, Cambridge, MA, pp. 15–118.

Kullback, S. (1959). Information Theory and Statistics, J. Wiley & Sons, Inc., London.

Malvestuto, F. & Pourabbas, E. (2004). Customized answers to summary queries via aggregate views, *Proceedings of 16th International Conference on Scientific and Statistical Database Management (SSDBM),* Santorini Island, Greece, IEEE Computer Society, pp. 193–202.

Malvestuto, F. & Pourabbas, E. (2005). Local computation of answers to table queries on summary databases, *Proceedings of 17th International Conference on Scientific and Statistical Database Management (SSDBM),* Santa Barbara, CA, USA, pp. 263–270.

Malvestuto, M. F. (1993). A universal-scheme approach to statistical databases containing homogeneous summary tables, *ACM Transactions on Database Systems* **18**(4): 678–708.

Palpanas, D., Koudas, N. & Mendelson, A. (2005). Susing datacube aggregates for approximate querying and deviation detection, *IEEE Transactions on KNowledge and Data Engineering* **17**(11): 1465–1477.

Pourabbas, E. & Shoshani, A. (2007). Efficient estimation of joint queries from multipe olap databases, *ACM Transactions on Database Systems* **32**(1): 1–43.

APPENDIX

A. THE LINEAR INDIRECT ESTIMATION

The main idea of such an approach is to use data from surveys of variables of interest at the national or regional level, and to obtain estimates at more geographically disaggregated levels 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 (Ghosh & Rao 1994). 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:

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

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 $\sum_i \hat{Y}(i)$ must be equal to $\sum_d Y(d)$ (Ghosh & Rao 1994).

B. AVERAGE RELATIVE ERROR

A method that is commonly used for measuring accuracy is the average relative error (ARE) (Ghosh & Rao 1994). 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.