AGGREGATE EVALUABILITY IN STATISTICAL DATABASES

F.M Malvestuto(†), M. Moscarini(‡)

(†) National Commission for Nuclear and Alternative Energy Sources, ENEA, Roma, Italy
(‡) Dipartimento di Matematica, Università di Roma "Tor Vergata", Roma, Italy

Abstract

Usually a statistical database contains many summary tables representing the distribution of the same statistical variable over the classes of as many partitions of a certain universe of objects. Existing query systems allow only queries on single tables. Indeed, in most cases additional queries can be evaluated by combining the information contained in similar tables in a suitable way.

In order to improve the responsiveness of the database and allow an integrated use of the stored information, we propose to inform the database system of the relationship among the partitions adopted in the tables. Such a relationship, called *intersection dependency*, states which classes of the partitions have a nonempty intersection and can be represented by a uniform multipartite hypergraph, called *intersection hypergraph*.

On the grounds of the algebraic properties of the intersection hypergraph and under the assumption of data additivity, we shall provide a characterization of evaluable queries, which allows us to define polynomial-time procedures both for testing evaluability and for evaluating queries.

1. Introduction

With the term *statistical database* we refer to "a numeric database containing statistics about classes of objects or individuals" [20] (social, economic, technological, environmental or demographic surveys are typical examples).

As usual, we shall assume that the unaggregated data from which the statistics have been computed is not available.

A statistical database is a collection of (*summary*) *tables*, each of which represents a distribution of a certain (*statistical*) *variable* (sometimes called "summary attribute" [14,20]) related to a given *universe* of objects or individuals, partitioned according to a set of (*category*) *attributes*, referred to as the *scheme* of the table.

Example 1. *Universe:* Soviet people in the year 1959. *Variable:* Population (1000 individuals). *Scheme:* {Sex, Schooling, Party-Membership} (the data is obtained by processing data from Bishop et al. [4]).

Table: Distribution of the soviet populatiion by schooling, sex and party (1000 individuals) 1959

Sex	Schooling	Party-Membership	
		Yes	No
Male	< 4	0	13670
	4-7	1217	20568
	8-10	2140	21135
	> 10	672	2341
Female	< 4	0	34730
	4-7	1071	18115
	8-10	2441	24109
	> 10	696	2424

□

For any subset S of the scheme of a table, all possible combinations of values of attributes in S define a partition of the underlying universe into *classes*.

Example 2. The classes of the partition defined by {S, Sc} in Example 1 are the groups of individuals qualified by the following conditions

S=Male ∧ Sc< 4
S=Male ∧ 4 ≤Sc≤ 7
S=Male ∧ 8 ≤Sc≤ 10
S=Male ∧ Sc> 10
S=Female ∧ Sc< 4
S=Female ∧ 4 ≤Sc≤ 7
S=Female ∧ 8 ≤Sc≤ 10
S=Female ∧ Sc> 10

□

Proceedings of the Fifteenth International Conference on Very Large Data Bases

Amsterdam, 1989

The partition defined by the whole scheme will be called a *base partition*; its classes correspond (one-to-one) to the entries in the table.

A typical query on a statistical database requires the evaluation of a variable for a subset (or *aggregate*) of objects in the underlying universe, qualified by a logical formula built up from specified values for given attributes by means of operators $\wedge$, $\vee$ and $\neg$. An aggregate is *evaluable* if the corresponding value to be assigned to the variable is uniquely determined by the data stored in the database (note that even in the case that an aggregate is not evaluable, a query can be answered in an "approximate" way [11]).

In literature [6,7,9,10,15,16,18,19] only logical formulas with attributes from a single table scheme have been considered. This means that every aggregate is the union of classes of a base partition, that is, an element of the set field generated by that base partition.

Example 3. Referring to the table in Example 1, a user can ask for the population of the aggregate specified by the following formula

$$\text{S} = \text{Female} \wedge \text{PM} = \text{Yes} \wedge (4 \leq \text{Sc} \leq 7 \vee 8 \leq \text{Sc} \leq 10)$$

□

If the data is *additive under aggregation* [8], the evaluation of a variable for an aggregate can be carried out by summing the entries corresponding to the classes contained in that aggregate.

Example 3 (continued). The answer to the query in Example 3 is

$$\text{Population}(x) + \text{Population}(y) = 1071 + 2441 = 3512 \text{ (1000 individuals)}$$

where

x: S = Female $\wedge$ PM = Yes $\wedge$ $4 \leq \text{Sc} \leq 7$

y: S = Female $\wedge$ PM = Yes $\wedge$ $8 \leq \text{Sc} \leq 10$ □

Indeed, a statistical database often contains "similar" tables, that is, distributions of the same variable over different partitions of the same universe of objects.

Similar tables may occur in the following typical cases:

- an organization ("data source") collects data on the objects of a given universe and then creates a database (called "abstract" in [17] and "partitioned database" in [5]) that is a collection of simple statistics;
- an organization creates a statistical database by putting together tables produced by different statistical agencies that refer to a common data source, but use different classification criteria.

Knowledge of relationships between attributes belonging to schemes of similar tables allows for additional aggregates to be evaluated [12,13]. The following example illustrates this statement in the case in which the *intersection dependency* between classes of the base partitions of similar tables is known, that is, we know which classes of the base partitions have a nonempty intersection.

Example 4. A statistical database contains the two following similar tables reporting the distributions of the employees of a research institute by department and by educational qualification. □

Table: Distribution of the employees by department

Department	Number
Information-Systems	30
Research	50
Administration	20

Table. Distribution of the employees by educational qualification

Educational-Qualification	Number
Ph.D.	15
High-School	20
Degree	65

Figure 1. Two similar tables.

Example 4 (continued). Suppose that we know the intersection dependency between the classes of the base partitions in Figure 1. We can represent this information with the bipartite graph in Figure 2, in which an edge represents a pair of nondisjoint classes of the two base partitions.

In this case, it is possible to evaluate aggregates which cannot be evaluated in traditional systems. An example of such an aggregate is

$$\text{D} = \text{Information-Systems} \wedge (\text{E} = \text{High-School} \vee \text{E} = \text{Degree}) \vee (\text{D} = \text{Research} \wedge \text{E} = \text{Degree})$$

In fact, on examining the graph in Figure 2 we note that the aggregate qualified by the formula above refers to the employees who work in the Information-Systems or Research departments and are not Ph.Doctors. Consequently, the number of qualified employees is

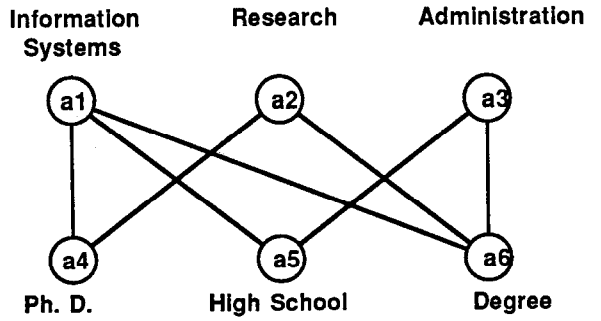

Figure 2. An intersection (hyper)graph

Number(a_1) + Number(a_2) - Number(a_4) = 30 + 50 - 15 = 65
where
a_1: D = Information-Systems
a_2: D = Research
a_4: E = Ph.D. □

The edges of the bipartite graph define the classes of a partition of the underlying universe that is finer than either of the two base partitions. Such a partition will be called *meet partition.*

Using formulas such as the one in the example (built up with attributes from the schemes of similar tables) we can express all the aggregates in the set field generated by the meet partition So, the following problem arises:

Problem. Is each aggregate in the set field generated by the meet partition evaluable ?

In this paper we approach this problem under the assumption that the data stored in similar tables is additive and that the underlying intersection dependency is known perfectly. In the general case in which the statistical database contains k similar tables, the intersection dependency can be represented by a uniform k-partite hypergraph [1] to be called *intersection hypergraph*, whose vertex set is in a one-to-one correspondence with the set obtained by union of the k base partitions, and whose edge set is in a one-to-one correspondence with the meet partition.

The paper is organized as follows. Section 2 contains the formal definition of an evaluable aggregate and the formulation of the problem of evaluability in the framework of linear equation systems. Section 3 contains an algebraic characterization of evaluable aggregates which leads to a quadratic procedure for testing evaluability, stated in Section 4. In Section 5 we shall show that evaluating an (evaluable) aggregate requires solving an equation system. Section 6 is dedicated to an illustrative example. Section 7 contains some basic properties of evaluable aggregates. In Section 8 we conclude by pointing out the advantages of an "informed" query system.

2. A formal framework

Let A be a partition of a *universe* Ω in *classes* of objects and A^+ be the *set field* generated by A, i.e., the family of subsets of Ω which are obtained as the union of a subset of classes of A. We call *aggregate* any element of A^+. A^+ turns out to be closed under set union, intersection and complementation.

Let F be a real-valued set function defined on a set field A^+; F is *additive* if for any two disjoint aggregates x and y in A^+, we have

$$F(x \cup y) = F(x) + F(y).$$

Observe that any additive function F on a set field A^+ is completely specified by the set of values $\{F(a) \mid a \in A\}$; in fact, by virtue of the additivity, for every $x \in A^+$ we have

$$F(x) = \sum_{\substack{a \in A \\ a \subseteq x}} F(a). \qquad (1)$$

Consider now k partitions $A_1, \ldots, A_k$ of Ω. The *meet partition* $E = \{e_1, e_2, \ldots, e_m\}$ of $A_1, \ldots, A_k$ is the partition defined as the greatest lower bound of $\{A_1, \ldots, A_k\}$ [2], i.e., the partition of Ω whose classes are given by the nonempty intersections $a_1 \cap \cdots \cap a_k$ with $a_i \in A_i$ $(i = 1, \ldots, k)$.

The set field E^+ generated by the meet partition E is the least set field containing $A_1^+, \ldots, A_k^+$ [13].

Given the partitions $A_1, \ldots, A_k$ of Ω, let $F_1, \ldots, F_k$ be additive functions on the set fields $A_1^+, \ldots, A_k^+$, respectively. By $\mathcal{F}$ we denote the family of additive functions F on the set field E^+ which satisfy

$$F(a_i) = F_i(a_i) \quad \text{for all} \quad a_i \in A_i \quad (i = 1, \ldots, k). \qquad (2)$$

By (1) the condition (2) can be written as

$$\sum_{\substack{e_h \in E \\ e_h \subseteq a_i}} F(e_h) = F_i(a_i) \quad \text{for all} \quad a_i \in A_i \quad (i = 1, \ldots, k). \qquad (3)$$

This is a system of $n = |A_1| + \cdots + |A_k|$ equations with m unknowns $F(e_1), F(e_2), \ldots, F(e_m)$; we call *partitioned homogeneous system* the homogeneous system associated with (3).

If $\mathcal{F} \neq \emptyset$, we say that the set of additive functions $F_1, \ldots, F_k$ is consistent.

An aggregate x in E^+ is *evaluable* with respect to a consistent set of functions $F_1, \ldots, F_k$ if either $|\mathcal{F}| = 1$ or, for every pair of functions F and F' in $\mathcal{F}$ we have $F(x) = F'(x)$.

In what follows, we assume that a collection $A_1, \ldots, A_k$ of distinct partitions of a universe Ω, their meet partition $E = \{e_1, \ldots, e_m\}$ and a consistent set $\{F_1, \ldots, F_k\}$ of additive functions be given.

We conclude this section with some remarks relating the formal notions introduced in this section to the concepts discussed in the Introduction.

- $A_1, \ldots, A_k$ can be interpreted as the base partitions of k similar tables contained in the database;
- the values $F_i(a_i)$ for each $a_i \in A_i$ are the entries of the i-th table;
- the additivity of the functions $F_1, \ldots, F_k$ formalizes the intuitive concept of "additive data under aggregation" mentioned in the Introduction, which applies to a large number of statistical variables (count data, probability data, measurement data, etc.);
- as discussed in the Introduction, similar tables are assumed to come from a unique data source; this ensures the consistency of the set of functions $\{F_1, \ldots, F_k\}$ and, hence, guarantees that it is meaningful to use the k tables in an "integrated" way;
  the classes of the meet partition are in a one-to-one correspondence with the edges of the intersection hypergraph;
  the coefficient matrix of system (3) (as well as the coefficient matrix of the partitioned homogeneous system) is the incidence matrix of the intersection hypergraph.

3. Characterization of evaluable aggregates

To every aggregate x in E^+ we shall associate an m-tuple $\mathbf{x} = [x(1), \ldots, x(m)]$, where $x(h)$ is equal to 1 if class e_h is included in x and equal to 0 otherwise. We call $\mathbf{x}$ the *representative vector* of the aggregate x. Observe that the rows of the coefficient matrix $\mathbf{A}$ of the partitioned homogeneous system are the representative vectors of the classes in the partitions $A_1, \ldots, A_k$. We call *row space* the vector space $\mathcal{R}$ spanned by the rows of the matrix $\mathbf{A}$.

Analogously, we represent a function F on E^+ by the m-tuple $\mathbf{f} = [f(1), \ldots, f(m)]$ where $f(h) = F(e_h)$ for all e_h in E. We call $\mathbf{f}$ the *representative vector* of the function F. We call *solution space* the vector space $\mathcal{S}$ of the solutions of the partitioned homogeneous system.

Finally, by $(\mathbf{x}, \mathbf{y})$ we denote the scalar product of two m-tuples $\mathbf{x}$ and $\mathbf{y}$.

Thus, by additivity for every F in $\mathcal{F}$ and for every x in E^+ we have

$$F(x) = \sum_{\substack{e_h \in E \\ e_h \subseteq x}} F(e_h) = \sum_h x(h) f(h) = (\mathbf{x}, \mathbf{f}).$$

Consequently, when $|\mathcal{F}| > 1$, an aggregate x is evaluable if and only if for every F and F' in $\mathcal{F}$

$$(\mathbf{x}, \mathbf{f}) = (\mathbf{x}, \mathbf{f}'). \tag{4}$$

Lemma 1. *The solution space $\mathcal{S}$ and the row space $\mathcal{R}$ are complementary subspaces of $\Re^m$*

Proof. By definition, $\mathcal{S}$ is the set of all the m-tuples that are orthogonal to the rows of the coefficient matrix of the partitioned homogeneous system; hence, $\mathcal{S}$ and $\mathcal{R}$ are complementary [3]. □

Recalling that the dimension of a vector space is the maximum number of linearly independent vectors, we have that the dimension of the vector space $\mathcal{R}$ is equal to the rank r of the matrix $\mathbf{A}$ and, by Lemma 1, that the dimension of the vector space $\mathcal{S}$ is $m - r$.

Lemma 2. *An aggregate is evaluable if and only if its representative vector is orthogonal to the solution space $\mathcal{S}$.*

Proof. If $|\mathcal{F}| = 1$ then system (3) has a unique solution; consequently, the partitioned homogeneous system also has a unique solution, namely the zero vector, which is orthogonal to every vector.

If $|\mathcal{F}| > 1$, let $\mathbf{f}$ be a solution of (3). The set of the solutions of (3) coincides with the set of vectors $\mathbf{f}'$ such that

$$\mathbf{f}' = \mathbf{f} + \mathbf{s}$$

where $\mathbf{s} \in \mathcal{S}$.

Therefore, by exploiting the bilinearity of the scalar product, for every x in E^+ we have:

$$(\mathbf{x}, \mathbf{s}) = (\mathbf{x}, \mathbf{f}' - \mathbf{f}) = (\mathbf{x}, \mathbf{f}') - (\mathbf{x}, \mathbf{f})$$

and by (4) the lemma is proved. □

Recalling that the coefficient matrix of the partitioned homogeneous system coincides with the incidence matrix of the intersection hypergraph, from Lemma 1 and Lemma 2 we can derive the following:

Theorem 1. *An aggregate is evaluable if and only if its representative vector is linearly dependent on the*

a_4: E = Ph.D.
a_5: E = High-School
a_6: E = Degree

The meet partition E of A_1 and A_2 is formed by 7 classes corresponding to the edges of the bipartite intersection hypergraph given in Figure 2:

$$e_1 = a_1 \cap a_4$$
$$e_2 = a_1 \cap a_5$$
$$e_3 = a_1 \cap a_6$$
$$e_4 = a_2 \cap a_4$$
$$e_5 = a_2 \cap a_6$$
$$e_6 = a_3 \cap a_5$$
$$e_7 = a_3 \cap a_6$$

The incidence matrix **A** of the intersection hypergraph is:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

The reduced matrix $\mathbf{A}'$ is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The partitioned homogeneous system in reduced echelon form looks like the following

$$\begin{aligned} s(1) \qquad\qquad -s(5) \qquad &= 0 \\ s(2) \qquad\qquad -s(7) &= 0 \\ s(3) \qquad +s(5)+s(7) &= 0 \\ s(4) \qquad +s(5) \qquad &= 0 \\ s(6) \qquad +s(7) &= 0 \end{aligned}$$

The corresponding basis of the solution space S is given by the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$, the former of which is obtained by setting $s(5) = 1$ and $s(7) = 0$ and the latter by setting $s(5) = 0$ and $s(7) = 1$:

$$\mathbf{s}_1 = [1\ 0\ -1\ -1\ 1\ 0\ 0] \quad \text{and} \quad \mathbf{s}_2 = [0\ 1\ -1\ 0\ 0\ -1\ 1]$$

Consider now the two aggregates $x = e_2 \cup e_3 \cup e_5$ and $y = e_1 \cup e_5$ with representative vectors

$$\mathbf{x} = [0\ 1\ 1\ 0\ 1\ 0\ 0] \quad \text{and} \quad \mathbf{y} = [1\ 0\ 0\ 0\ 1\ 0\ 0]$$

Note that aggregate $\mathbf{x}$ is the aggregate considered in Example 4.

It is immediately seen that $(\mathbf{x},\mathbf{s}_1) = (\mathbf{x},\mathbf{s}_2) = 0$. So, x is an evaluable aggregate. But, $(\mathbf{y},\mathbf{s}_1) \neq 0$ and, therefore, y is not evaluable.

In order to express $\mathbf{x}$ as a linear combination of the rows of $\mathbf{A}$, take as a basis of row space $\mathcal{R}$ the set $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5\}$ of rows of $\mathbf{A}$ corresponding to the nonnull rows of the reduced matrix $\mathbf{A}'$. The coefficients λ_q of the required combination can be determined by solving the system

$$\mathbf{x} = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \lambda_3\mathbf{a}_3 + \lambda_4\mathbf{a}_4 + \lambda_5\mathbf{a}_5$$

by successive elimination of the unknowns; an admissible solution is given by:

$$\lambda_1 = 1$$
$$\lambda_2 = 1$$
$$\lambda_3 = 0$$
$$\lambda_4 = -1$$
$$\lambda_5 = 0$$

and, therefore, we have

$$\mathbf{x} = \mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_4.$$

and

$$(\mathbf{x},\mathbf{f}) = F_1(a_1) + F_1(a_2) - F_2(a_4).$$

7. Properties of evaluable aggregates

In this section we provide some basic properties of evaluable aggregates.

Proposition. *Let x and y be evaluable aggregates in E^+. We have:*

(i) *(disjoint-union property): if $x \cap y = \emptyset$, then $x \cup y$ is evaluable*

(ii) *(proper-difference property): if $x \supseteq y$, then $x - y$ is evaluable*

(iii) *(complement property): $x' = \Omega - x$ is evaluable.*

Proof. The statements can be easily proved by additivity. In particular for every F in $\mathcal{F}$ we have:

- in case (i), $F(x \cup y) = F(x) + F(y)$
- in case (ii), $F(x - y) = F(x) - F(y)$
- in case (iii), $F(x') = F(\Omega) - F(x) = \sum_{a_i \in A_i} F_i(a_i) - F(x)$ for any A_i. □

rows of the incidence matrix of the intersection hypergraph.

Proof. In fact, by Lemma 1 every vector orthogonal to the solution space S belongs to the row space $\mathcal{R}$. So, the theorem follows from Lemma 2. □

Corollary 1. *Each aggregate in E^+ is evaluable if and only if the dimension of the row space $\mathcal{R}$ equals the cardinality of the meet partition E.*

Corollary 2. *Each aggregate in E^+ is evaluable if and only if the solution space S contains only the zero vector* **0**.

4. Testing evaluability

Lemma 2 allows us to polynomially test the evaluability of an aggregate x in E^+ once a basis for S is given (recall that a basis of S is a set of $m - r$ independent vectors).

It is well-known [3] that a basis for S can be determined by transforming the partitioned homogeneous system in the so-called *reduced echelon form* [3], whose solution space coincides with S. If the matrix **A** has rank r, we have

$$\begin{aligned} s(1) &+ c_{1,r+1}s(r+1) + \cdots + c_{1,m}s(m) = 0 \\ s(2) &+ c_{2,r+1}s(r+1) + \cdots + c_{2,m}s(m) = 0 \\ &\vdots \\ s(r) &+ c_{r,r+1}s(r+1) + \cdots + c_{r,m}s(m) = 0 \end{aligned}$$

This can be done [3] by first reducing the matrix **A** to a row-equivalent matrix **A′** by repeatedly applying the following row operations:

(i) replace row $\mathbf{a}_i$ by $c\mathbf{a}_i$ for any scalar $c \neq 0$

(ii) replace row $\mathbf{a}_i$ by $\mathbf{a}_i + d\mathbf{a}_j$ for any $j \neq i$ and any scalar $d \neq 0$

and, then, rearranging the rows of **A′**.

Therefore, the task of transforming the partitioned homogeneous system into its reduced echelon form can be accomplished with $O(mn^2)$ elementary operations.

A basis $\{\mathbf{s}_1, \ldots, \mathbf{s}_{m-r}\}$ of S can be obtained by taking the $m - r$ **s**-solutions resulting from setting one of the parameters $s(h)$ $(h = r+1, \ldots, m)$ equal to 1 and the remaining parameters equal to 0 . That is,

$$\begin{aligned} \mathbf{s}_1 &= [\, -c_{1,r+1}, \ldots, -c_{r,r+1}, 1, 0, \ldots, 0] \\ \mathbf{s}_2 &= [\, -c_{1,r+2}, \ldots, -c_{r,r+2}, 0, 1, \ldots, 0] \\ &\vdots \\ \mathbf{s}_{m-r} &= [\, -c_{1,m}, \ldots, -c_{r,m}, 0, 0, \ldots, 1] \end{aligned}$$

At this point in order to decide whether x is evaluable or not, it is sufficient to verify that the scalar product of its representative vector **x** with each basis vector $\mathbf{s}_j$ $(j = 1, \ldots, m-r)$ of the solution space S vanishes.

It should be noticed that since a basis for S can be determined a priori on the grounds of the intersection dependency, the test for evaluability can be carried out without accessing the database.

From the foregoing discussion the following theorem follows.

Theorem 2. *Testing evaluability requires $O(m^2)$ elementary operations.*

5. Computing answers to evaluable queries

To answer a query related to an evaluable aggregate x, the database system needs to determine the coefficients of the linear combination of the rows of **A** which provides **x**.

Let $\mathbf{a}_1, \ldots, \mathbf{a}_r$ be the rows of **A** which give rise to the nonnull rows of the reduced matrix **A′** mentioned in the previous section. The set $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ is a basis of the vector space $\mathcal{R}$ and, therefore, for each evaluable aggregate x we have

$$\mathbf{x} = \sum_{q=1}^{r} \lambda_q \mathbf{a}_q \qquad (7)$$

which is a system of m equations with r unknowns, namely, $\lambda_1, \ldots, \lambda_r$, and can be solved using standard methods.

Once the coefficients λ_q have been determined, the answer to the query is given by

$$F(x) = (\mathbf{x}, \mathbf{f}) = \sum_{q=1}^{r} \lambda_q (\mathbf{a}_q, \mathbf{f}) = \sum_{q=1}^{r} \lambda_q F_{i_q}(a_q)$$

where a_q is the class in the partition A_{i_q} with representative vector $\mathbf{a}_q$ and F_{i_q} is the additive function on the set field generated by A_{i_q}.

6. An example

Consider the two base partitions $A_1 = \{a_1, a_2, a_3\}$ and $A_2 = \{a_4, a_5, a_6\}$ induced by the attributes Department and Educational-Qualification in Example 4, where

a_1: D = Information-Systems

a_2: D = Research

a_3: D = Administration

Let us consider, now the minimal evaluable aggregates, that is, those evaluable aggregates which contain no proper subsets which are evaluable

Theorem 3. *Every evaluable aggregate is a disjoint union of minimal evaluable aggregates.*

Proof. Let x be an evaluable aggregate which contains a minimal evaluable aggregate y. Then, by the proper-difference property the aggregate

$$z = x - y$$

is evaluable, too. So x is the disjoint union of y and z. If z is also minimal, then the theorem is proved; otherwise, repeat the same line of reasoning made for x to z. By finite induction, we prove that x is a disjoint union of minimal evaluable aggregates. □

8. Concluding remarks

In this paper we approached the problem of the evaluability of aggregates in a statistical database containing similar tables related by an intersection dependency.

We provided a characterization of evaluable aggregates. Our characterization allows us to define procedures both for testing evaluability and for evaluating queries.

These results are useful in designing an "informed" query system for statistical databases which promotes an integrated use of stored information. Such an "informed" query system allows the user to formulate a query involving attributes from several similar tables as if they were all contained in a single table ("universal" table), that is without knowing that these attributes have been extracted from different tables; for example, a user queries the database in Example 4 using the "universal" scheme {Department, Educational-Qualification}.

References

[1] C. Berge, *Graphs and Hypergraphs*. North Holland, 1973.

[2] G. Birkhoff, *Lattice Theory*. American Mathematical Society, 1967.

[3] G. Birkhoff and S. McLane, *A Survey of Modern Algebra*. Mc Millan, 1962.

[4] Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, 1975.

[5] R. Brooks. M. Blattner, Z. Pawlak and E. Barret, "Using Partitioned Databases For Statistical Data Analysis", *AFIPS Conference Proc.* (1981). 453-457.

[6] E. Fortunato, M. Rafanelli, F.L. Ricci and A. Sebastio, "An Algebra for Statistical Data", *Proc. III Int. Workshop on "Statistical & Scientific Database Management"* (1986), 122-134.

[7] S.P. Ghosh, "Statistical Relational Tables for Statistical Database Management", **IEEE Software Engineering 12** (1986), 1106-1116.

[8] S.P. Ghosh, "Statistics Metadata", **Encyclopedia of Statistical Sciences 8** (by Kotz-Johnson). J. Wiley & Sons, 1988.

[9] G. Hebrail, "A Model of Summaries for Very Large Databases", *Proc. III Int. Workshop on "Statistical & Scientific Database Management"* (1986), 143-151.

[10] A. Klug, "Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions", **JACM 29** (1982), 699-717.

[11] F M. Malvestuto, "Answering Queries in Categorical Data Bases", *Proc. ACM 1987 Symp. on "Principles of Database Systems"*, 82-89.

[12] F.M. Malvestuto, "The Derivation Problem for Summary Data", *Proc. ACM SIGMOD 1988 Conf. on "Management of Data"*, 87-96.

[13] F.M. Malvestuto and C. Zuffada, " The Classification Problem with Semantically Heterogeneous Data", **Lecture Notes in Computer Science 339** (M. Rafanelli, J.C. Klensin and P. Svensson eds.), Springer Verlag 1989, 157-176.

[14] F. Olken, D. Rotem, A. Shoshani and H. Wong, "Scientific and Statistical Data Management Research at LBL", *Proc. III Int. Workshop on "Statistical & Scientific Database Management"* (1986), 1-20.

[15] Z.M. Özsoyoğlu and G. Özsoyoğlu, "An Extension of Relational Algebra for Summary Tables", *Proc. II Int. Workshop on "Statistical Database Management"* (1983), 202-211.

[16] G. Özsoyoğlu, Z.M. Özsoyoğlu and V. Matos, "Extending Relational Algebra and Relational Calculus with Set-valued Attributes and Aggregate Functions", **ACM Trans. on Database Systems 12** (1987), 566-592.

[17] N.C. Rowe, "Antisampling for Estimation: an Overview", **IEEE Trans. on Software Engineering 11** (1985), 1081-1091

[18] H. Sato, "Handling Summary Information in a Database: Derivability", *Proc. ACM-SIGMOD 1981 Conf. on "Management of Data"*, 98-107.

[19] H. Sato, "Fundamental Concepts of Social-Regional Summary Data and Inferences in their Databases", (thesis) Japan Economic Planning Agency, Tokyo 1982.

[20] A. Shoshani, "Statistical Databases: Characteristics, Problems and Some Solutions", *Proc. VIII Int. Conf. on "Very Large Data Bases"* (1982), 208-222.