Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2002/09

The temporal dimension in data bases

Cristina Martelli

Università degli Studi
di Firenze

Applied Statistics

The Temporal Dimension in Data Bases

Cristina Martelli
Department of Statistics
University of Florence- Italy

Abstract

The advantages brought by the proper management of information in terms of quality, integrity, and robustness in data manipulation is widely known. This is particularly true for statisticians that must grapple with multidimensionality in data, complex semantics in queries, and large quantities of data needing to be elaborated. In this context, an appropriate approach to temporal information is especially important: dynamic analyses are involved in many statistical fields, and information systems must be adequate to face the complexity of temporal logic. In particular, the structure of statistical sources must be optimised for a complete exploitation of their informative potentialities and in order to make them linkable to other sources, with the respect to their temporal coherence.
The purpose of this paper is to reflect upon the potentialities that a correct approach to information management can bring to statistical analysis, with particular attention to the problem of temporal data storing.

1. INTRODUCTION

Reflection on the conceptual link between data and information has been deeply developed in the last two decades; one of the main outputs in this field has been the concept of *database*, not simply in its technical and instrumental sense, but, more generally, as a conceptual approach to the problem of information management.
In the classical *database approach*, in fact, one priority is to succeed in formulating a high level, logical description of the collected information structure: data are described through a *data conceptual model*, and this model should be as independent as possible from the physical representation of data (the files), in order to guarantee that modifications made to the physical representation do not involve modifications to the logical one. As a consequence of this approach, it is possible to get a separation between logical and physical aspects that is often referred to as *data independence.* (Ullmann,1989; Atzeni De Antonellis,1993)
Taking this methodology as a starting point, it is possible to achieve many results: first of all, different programs (that are the expression and consequence of different organisational and informational needs) may access and modify the same database and share common data, thereby reducing the inconsistencies and redundancies among the representations of the same data in different programs.
Another important consequence is represented by the possibility of using query languages for direct access to the data. These querying tools enable the user to directly organise the data in the most suitable way to his or her own informational needs, without having to make any investment in programming.
Finally, the possibility of bringing together various informational sources is greatly enhanced by the structuring of the sources, by a clear outline and individuation of all the informational actors present in the structure, and by the specification of the existing relations between them. Independence between data and programs is provided by software systems, called Database Management Systems (DBMSs), which give common and controlled means for accessing and modifying data, along with an integrated set of services, which include support for security, integrity and reliability of the data.
In recent years, broad consideration of the possibilities offered by the use of databases in the field of statistics has been developed. Above all, reflection has focused on what is meant by the term “statistical database.”
There are two ways one can interpret the meaning of statistical database. The first, and more limited view sees this merely as an archive of aggregated statistical data (Rafanelli, Klensin and Svenson.,1989, Michalewicz, 1990, Hinterberger, 1992). On the other hand, different view does not confine itself to a typology of data but to the way the data is used and defines the statistical database as a filing system able to represent information in such a way that is suitable for statistical analysis. Of the two, we find ourselves in line with the second, more broad interpretation.
The advantages that stem from the use of database for statistical sources structuring are of different types. Above all, from a more instrumental point of view, one recalls the possibility of having integrated workplaces made up of databases for the data management and packages for statistical analysis. In this type of organisation, packages query the database directly, extract the view most suitable to the particular necessities at hand, and carry out the elaboration, perhaps downloading the results in the files, without the need to duplicate or to write additional programs.
From the methodological point of view we recall, above all, the benefits given by a unified management of the surveys in which, on the basis of a single conceptual model, the questionnaire and the file that must contain the collected data are structured. In practice, this approach is not yet widely used; questionnaires are only conceived from the starting point of variables that are intended to be gathered and by the most suitable sequence for questions. The result is that the archives are disconnected from their conceptual model and the data are not directly queriable (Martelli and Casini, 1999).
Another important point of interest for a statistical user is that, because modern information management systems allow for the efficient storage and retrieval of information, data and also statistical sources have grown. There is, in fact, a large amount of administrative data that were not originally collected for statistical purposes, that nevertheless have tremendous potential when directed to that end. For example, routine patient records in hospitals can be used for statistical cause and effect studies and business transactions can be statistically analysed for econometric models or for policy settings.

In this sense it is important to consider the experience of several countries[1] that have substituted most of the traditional statistical sources, like surveys and censuses, with an efficient handling of administrative registers, seen and used as a complex and integrated informational system (Buzzigoli and Martelli, 1999 (a) and (b)).
The point is to succeed to exploit, by the statistical point of view, the funds of administrative data that accumulate in the informational systems of enterprises and institutions. The recent considerations that have been gathered pertaining to the methods and instruments for the data warehouse (Kimball, 1996; Inmon, 1995) have brought to light that an efficient statistical use of these resources obliges to reflect on possibilities and limits of the traditional modelling approaches used in the productive environment.
A third fundamental ground for interest lies in the consequences of the use of database type approaches on the quality of information that is produced. The availability of query languages that are directly utilisable, allows, in fact, for the efficient set-up of protocols for quality control of data aimed at searching for errors, missings and incoherences. It is also worth remembering that there is the possibility of enriching the system with metainformation that guides the use and maintenance of the source itself (McClean, O.I. Grossman W., Froeschl K.A., 1998).
In the light of these early considerations it is easy to understand that there are many reasons that justify statisticians' interest in methods and techniques for the management of databases. However, despite recent advances in research on methodologies and techniques for statistical information sources organisation, and while elsewhere the concept of data independence is well understood and applied, it is normal for a statistician to have to face huge amounts of destructured data that can be used only after taking great lengths to match the elaborative models and procedures.
A statistical database in day to day work life is very often nothing but a flat file from which the desired variables are selected programming a series of successive selection and recodification steps, thereby forfeiting the chance to benefit from the advantages we have just mentioned.

2. TIME IN STATISTICAL DATABASES

Statistical sources, contrary to those used in production, are substantially historical; in non-statistical applications, in fact, it is usually important to know only the most up-to-date level of information, and frequently the user is not at all interested in knowing the history of the values assumed by a certain variable. This is why most conventional databases represent reality only at the current time; the current contents of a database can be viewed, in this sense, as a *snapshot* of the real world at a single instant of time. As the real world changes, new values are incorporated into a database by replacing the old values.
Statisticians have different needs: the proper management of the temporal component of a source is, in fact, of fundamental importance for them, as all social, economic, and productive activities, that traditionally are the statisticians' object of research and analysis, always occur in a temporal context. In this regard, it is interesting to recall all the fields focused on the study of evolutionary behaviours or on the evaluation of links between causes and effects; another important point, next to the problems of temporal series archiving optimisation, is linked to the predisposition of integrated sources (for example, for micro-macro analyses) in which the aim is to bring together, in a context that is temporally homogenous and therefore queriable as a *unicum*, contexts described with different temporal metrics.
Other fundamental applications are referred to all those problems connected to the measurement and analysis of source's quality. For example, in order to evaluate the construction process of an indicator, one would like to archive and efficiently use all the events that mark the history of its making, or to have a proper metainformation apparatus that assists in the management of codes and definitions evolution.
The ascertainment of the lack of a correct modelling of the source, that limits and penalises the way in which it is used, is already pushing toward the reorganisation and restoration of archives that were originally laid out in periods that were not equipped for this kind of approach.
In this regard, one thinks of the present experiences going on to recuperate the longitudinal dimension of survey data that were at the time of the survey filed in a way aimed only at cross-section use (Elder G.H., Pavalko E.K, Clipp E.C., 1993).

---

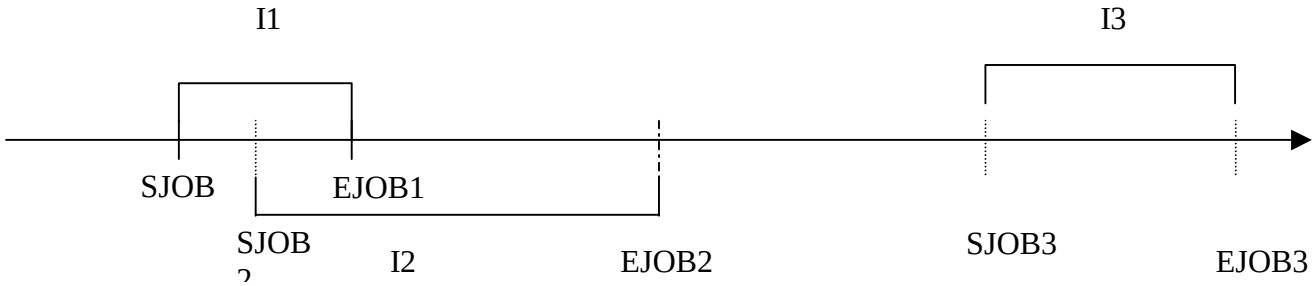
[1]By this point of view is peculiar and particularly interesting the danish experience; the Danish statistical system is based since twenty years on a system of linkable (via the personal identity code) administrative registers. Actually this system has completely substituted the censuses. Danmarks Statistik, *Personstatistik i Danmark: et Registerbaseret Statstiksystme,* 1994. C.Martelli,1995)

An information management system useful for statistical analysis should therefore furnish a whole series of uses/benefits and facilities for modelling and managing the dynamic component of the gathered data. A temporal database preserves the complete history of objects by retaining their previous values, and, in most cases, what constitutes a replacement or modification of an old value in a conventional database becomes an insertion of a new fact in a temporal database. (Tansel, Clifford, Gadia, Jajodia, Segev, Snodgrass,, 1993)
Considerable research activity has been directed recently to the study of time in databases.[2]
Consideration has focused on two principal points: first, there was the need to understand what original modelling instances had to be considered in order to obtain a complete representation of the temporal component that had to be managed by the database. It was also necessary to reflect on the instrumental and technical aspects of the solution that was intended to be produced.
From the methodological point of view, discussion obviously began with the problem of inserting time into the conceptual model of the source. The role of the conceptual model in a database approach is, in fact, that of explicating the structures, operations and constraints of the information system. The structural component of a data model deals with objects and their relationships while the operational/behavioural component deals with their manipulation. The constraint component deals with rules for the integrity of the information structure.
The two most prominent models that have provided the basis for the development of conceptual models have been the entity relationship model (Chen 1976) and the object based model. The entity relationship model deals with the structural components and is founded on the notion of entity and relationship. The object model deals with both the structural and behavioural components and is founded on the notion of object, structure and behaviour.
Furthermore, a number of approaches include notions like event-condition-action that deal with the constraint component.
To introduce the notion of time in modelling activity different approaches have been proposed, starting from the extension of the semantic of pre-existing snapshots model in order to directly incorporate time. Other proposals have suggested basing new models on a snapshot model with time appearing as an additional attribute or suggesting to move in an independent direction and developing entirely new approaches.
Nowadays it is generally agreed that the next generation of temporal data models should be an extended model, rather than an extensible one, with respect to the current generation.
Whatever the approach, highest priority must be given to consideration of the types of time domains supported: some functionality, in fact, can be provided only if a particular type of time domain is supported. For instance, it is not possible to reason about time at different levels of granularity, if it is not possible to define temporal data at different granularity levels. Similarly, it is not possible to talk about temporal distances between events if no metric is supported (Montanari e Pernici,1993). The granularity of a given time element is the level of abstraction at which the element is defined. In a calendar based system, for instance, the day level of granularity has the day as the unit of time at the abstraction level; to handle multiple time granularities, such as. days vs. weeks, it is necessary to consider multiple interrelated temporal domains. An instant in a " higher-level" domain corresponds to a contiguous set of instants in another "lower-level" domain.
Time metric introduces the concept of distance in time or duration, depending on whether time points or time intervals are considered.
Temporal systems, in fact, can be based either on the primitive notion of time point or on the primitive notions of time interval.
The choice between a time-point or an interval approach is perhaps the first decision to take when modelling a source for dynamic statistical analysis. As with the dichotomy between time intervals and points, a dynamic system can be conceptualised in terms of events or in terms of states (Martelli, 1999(a)). In particular, an event happens in a particular instant (following some action in the real world); for modelling an event one needs a single instant of time, the instant in which the event takes place. For modelling an interval, two instants of time are required, the starting and the ending times. The history of an entity is made by sequences of states; the passage from one state to another is marked by events.
Consider, for instance, the problem of organising the data collected by a survey focused on the description of the respondent's work experiences. The respondent will probably answer in terms of dates: when he or she found each job and when these experiences ended.

---

[2]At the ARPA/NSF Workshop on Infrastructure for Temporal Database, held in Arlington,TX on June 1993 it has been remembered that temporal databases had a corpus nearing 800 papers, in the last ten years, but surprisingly, in spite of this substantial activity, we are still far from a wide usage in the community of user in general and statisticians in particular.

Graph 1



The time axis of graph n^1 represents the biography of a respondent who has experienced three different jobs: for each job, the respondent has given the starting and ending dates (SJOBn and EJOBn, respectively).
These dates are events (or, time points). Looking at this information from another point of view, the respondent has experienced different states: without job, with one job, with at least one job, with the second job, with the third job.
Depending on the type of analysis proposed for these data, the temporal data model can favour an event point of view or an instant point of view. The choice will depend, for instance, on the type of operators that are needed in the analysis.
This choice is not necessarily a dichotomy: there are in fact examples (Vilain, 1982) in which every relation type is explicitly modelled.
By the point of view of points and intervals mathematical structure, it is important to note that for both of them it is possible to apply an *ordering relation.* An ordering relation is in fact defined over time-point domains as well as over time-intervals domains; in point based systems, intervals are defined as pairs of time points, the lower and the upper ends of the interval. Ordering relations between intervals are expressed in terms of relations between their endpoints. In interval based systems points are defined as the beginning and ending of intervals and identify the “places” where intervals meet.
The most common orders studied in temporal databases are usually linear, and the assumption of linearity of time is at the basis of most temporal models: when time points are considered, time is said to be linear if the set of time points is totally ordered. Some alternatives have to be considered in the possible evolution of temporal data; for instance, circular time can be used to represent recurrent events; in this case no ordering relationship can be defined (Montanari e Pernici, 1993).
Another distinction that must be made is between relative and absolute times. With absolute times it is defined a location of a time point or a time interval on the time axis; in general they are associated with a temporal metric, such as calendar times and with an origin defined on the time axis. In this way a time point can be associated with a particular date and time intervals with a given interval on the calendar. When relative times are used, as they don’t have a precise location on the time axis it is interesting to define only their position relative to other times (points or intervals). This distinction is fundamental, for instance, for information systems that are designed to handle survival data.
As for the instrumental side of the temporal data management problem, one starts observing that commercial database management systems have traditionally shown themselves to be inadequate to support temporal statistical application; they were in fact designed to support transactions for business applications, and are often inefficient when using the large amount of data existing in statistical and scientific applications. Very often analyses are not carried out at the desired level of granularity because it is impractical to analyse the large amount of data that would consequently be generated.

3. TIME AND THE RELATIONAL MODEL

The relational approach (Codd,1970) is a way of thinking about databases, and it is one of the most popular and elegant ways for modelling reality. It couples a precise mathematical definition with a useful representation based on tables, and it has been formulated in order to respond to the requirement of data independence. Previous models, either hierarchical or network, included explicit reference (through pointers or links) at the logical level to the underlying physical level. Another important motivation supporting the adoption of the relational model was its flexibility with respect to a wide variety of possible operations, especially queries, that make it easier to implement for the availability of a formal and algebraic approach to the modelling problem.

The relational model makes use of a single structure to organise data: this structure is a variant of the mathematical concept of n-ary relation: a relational database is represented as a collection of tables where every table has a different and unique name in the database. A row in the table represents a relationship among sets of values, column headings contain distinct names, and for each column there is a set of possible values, called the *domain*. Observing this tabular representation of information, it is easy to recognise a strong correspondence between the concept of relation as outlined in mathematical set theory, and the table of this tabular representation of information.

The study of problems aimed at the elimination of anomalies leads to the definition of *normal forms* for relation schemes, that, when respected, assure in practice that no attribute can depend on any set of attributes that is not a key.

At present, the major point of interest for a statistician in the relational approach, beyond the obvious management advantages, is tied to the fact that most administrative sources are managed in a relational manner. Even at the national level, these archives are laid out in such a way as to be easily managed in tabular form, with these tables having links among themselves. In public administrations, for example, the introduction of unique personal and non-ambiguous keys that identify the same subject in different contexts has enhanced the possibility of evaluating the informational patrimonies of the single sectors, opening the way for the possibility of linkages and_integration. In other words, wherever an individual code (i.e., for citizens) is adopted, the individual, being identified by an non-temporal key, becomes an object in the relational sense of the term.

The repercussion for citizens' privacy, given by the fact that they live in a national relational type informational systems, is obvious and, in this vein, in the sphere of the different national contexts, suitable legislative instruments were made available that aimed at defining the legitimacy of the navigation in the sphere of these linkable filing systems.

As regards business statistical sources, the importance held by the relational approach is even more evident: in this case, in fact, informational systems are not only characterised by a relational structural model, but are also practically managed by Relational Data Base Management Systems, that have become the technical standard for the archival management of data in a very broad context.

In light of the importance held by the relational model in the source construction for statistical analysis, and considering the importance that the temporal dimension of information has for informational statistical systems, much research activity has been directed to making the relational approach be adequate to a time varying reality. In particular, research has been aimed at developing solutions consistent with existing commercial relational DBMSs that have been very widely used to model and manage snapshot databases.

The effect on the informational structure of an observer that observes the dynamic aspects of reality and records the data, produces a hierarchical effect on the organisation of the conceptual model (Martelli, 1999 (b)); when, in fact, one or more of the attributes of a table changes in time, the application of the relational point of view (via the first normal form that resolves the hierarchical structures) leads to a fragmentation of the description of the tuple. On the contrary, in the real world, even though properties of an object change in time, we think of it as the same object. In practice the insertion of time into a description of reality makes us lose the correspondence between the tuples and the different "actors" acting in the situation. So, the goal would be that a single tuple could capture the entire history of an object.

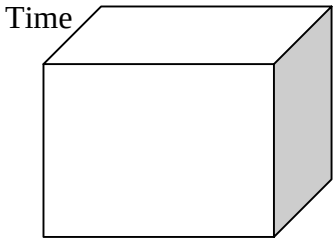
Let us offer an example to better clarify this problem: physicians faced with diagnostic and therapeutic decisions must reason about clinical features that change over time. For every patient the database will record sets of information repeated in time like chronology of disease symptoms, exacerbation, remissions, cures and failures. Thus, there is a hierarchy between the patient and the different temporal steps of his or her clinical story, and this structure results in practice in a fragmentation of the entity *patient* in the different temporal steps of his or her life history.

Reflection on the adaptation of the relational approach to the management of dynamic information has so faced, above all, the problem of the tuples fragmentation, focusing meanwhile, the formulation of criteria of normalization that would include a consideration of the temporal dimension.

Many solutions have been proposed by various authors to the problem presented by the combined goals of avoiding fragmentation and maintaining consistency with commercial relational databases and their query languages. Following the value-oriented nature of the relational model, some authors assume that, for applying the relational model to a time varying situation, an object must be identified by key time invariant attributes, and that all the non-key attributes in a tuple must have the same time domain.

This requirement is called *homogeneity* and it ensures that an instantaneous snapshot of the tuples does not contain any nulls.(Gadia and Vaishnav, 1985; Gadia 1988)

If the pace at which all the patients' attributes change in time is the same, the representation of the classical relational model tuple will result in a cube: some authors, in fact, propose a cubic view of a database to capture the time dimension (Ariav,1986). Here, time is added as a third dimension to the two dimensional flat tables of the relational data model.



At the representational level, relations are extended to include time attributes at the tuple level or at the attribute level. The primary reason for advocating attribute time stamping is that values of attributes within a relation vary at different rates. In the attribute stamping approaches, each tuple contains a history of values for each attribute. Consequently, such relations are not even in the first normal form.(Tansel, Clifford, Gadia., Jajodia, Segev, Snodgrass, 1993)
Other authors have reflected on the nature of "normalization", and have extended a definition of normalization to the time domain (Navathe and Ahmed, 1989) .
By definition, a relation is a set; hence its elements (tuples) must be independent of one another. In the relational database theory a tuple is the most fundamental unit of information: when a relation is normalized the different tuples become semantically independent. The goal of time normalization is getting tuples semantically independent of one another in order to avoid redundancies and anomalies in temporal updating and retrieval. To achieve this point, it is important to reflect on the conceptual notion of synchronicity, defining different types of synchronism among time varying attributes, and on the concept of temporal dependency, which is used to define the notion of time normalization.
A set of time varying attributes in a given relation is called synchronous if every time varying attribute can be uniformly associated with and be directly applied to the time stamps values in each tuple of the relation.
The presence of more than one synchronous equivalence class in a relation implies that some time varying attributes change in an asynchronous fashion; such asynchronies lead to the fragmentation of the lifespan information of a time varying attribute over several tuples and creates updates and retrieval anomalies.
The notion of temporal dependency and time normal form have been so proposed, which can successfully avoid retrieval and update anomalies and redundancies.
For a relation to be in a time normal form means that it not contain attributes that are not in the same synchronous class: a relation is in time normal form (*TNF*) if and only if it is in *BCNF (Boyce Codd Normal Form*)[3] and there exist no temporal dependencies among its non-key attributes.
It is always possible to decompose a relation, if a temporal dependency exists, into two or more time normalized relations by appropriately partitioning the attributes and merging the relevant time intervals. This decomposition satisfies the property that no two temporally dependent attributes remain in the same relation. In general, the set of time varying attributes in a relation can be partitioned into a minimum number of sets such that no two attributes within one subset are temporally dependent on each other.
The relational model is able to manage either an approach to time based on intervals or on points, that, as was mentioned in the preceding paragraph, are two of the possible ways in which time can be conceptualised.
In the case of a conceptualisation aimed at events, and therefore at points, the event will be an entity characterised by the date in which it happened. Information on states will be obtainable as queries stemming from this base. In the case where, instead, a conceptualisation for intervals is desired, these can be defined by their extremes, which are considered as start and stop point of the state.
While an event relational management presents no implementation problems, that of interval requires special attention.
As mentioned in the previous paragraph, the concept of an interval is quite general and has numerous application areas. Intervals and their properties are known from mathematics, but their mathematical definition

[3] "A relation scheme with some keys and some functional dependencies, is in Boyce-Codd normal form (BCNF) if for every dependency X--->Y (where Y is not a subset of X) the set of attributes X contains a key for the relation scheme" Atzeni, De Antonellis, 1993

and the set operations on them cannot be readily applied to relational snapshot databases. For example, an interval can be an uncountable infinite set, whereas in a database only a finite number of elements can be recorded. Furthermore, in mathematics, the set-union and set-difference operations of two intervals are not closed. A specific formalisation for their modelling has been proposed in the context of the Interval Extended Relational Model (Lorentzos and Johnson, 1987, 1988; Lorentzos, 1993) . It can effectively handle generic intervals, and it involves adding only two operations to the snapshot relational model, one to extract points from an interval and another that forms intervals from successive points**.**

4. NON-RELATIONAL APPROACHES

Even if most of the work in the area of temporal databases has been done in the context of the relational model, several recent studies have been directed toward going beyond relational data models. The increase in complexity of new applications such as the scientific and statistical databases has showed the limits of the relational model and has lead to research into next generation data models, including entity/relationship, object-oriented and deductive data models. These data models have been thought to capture the semantics of complex objects and treat time as a basic component rather than an additional attribute (Tansel, Clifford, Gadia., Jajodia, Segev, Snodgrass, 1993).

Moreover, as opposed to what it happens in informational systems aimed at management and production, statistical sources are essentially historical, and for them the problem of data updating is not essential. The relational approach allows, in fact, for the elimination of anomalies during the steps of data insertion and cancellation thanks to an exclusive dependence of the attributes on the key and by means of respect of the data non-redundancy condition. In the case in which historical data (and therefore stable) is being filed, the problem of management of updating and changes in data is hardly posed any more. In statistical archives the only changes are those that serve to correct errors. In light of these statistical sources characteristics falls, above all, the need to ensure the non-redundancy of data, typical of the relational approach; it is possible to choose redundant modelling approaches less efficient in the retrieval of a single instance but able to sustain richer and more complex query semantics. This type of layout is at the base of the so-called data warehouse approach, that theorises the separation of the statistical and analytical sphere(that uses statistical historical data) from the management sphere(for which is compulsory to work with non redundant data) by means of a duplication of information and a clear separation of the two environments, even at the level of conceptual modelling that remains relational only at the management level.

The comparison of the non-relational temporal data model with the relational temporal data model is analogous to the comparison between the snapshot version of the models. For example, a model such as the entity-relationship model has more semantics and is more user friendly than the relational model. The same is true for the temporal version of the object-oriented models.

New primitives to handle temporal data and complex types need to be incorporated into the query and specification languages of these models. Recently, there have been several research efforts in the area of temporal object-oriented data models (Rose and Segev, 1991, 1992). Reflection about the statistical database, in fact, has recognised the importance of object modelling for statistical information systems early on (Kim, 1990): the need for multidimensional data-modelling, typical of statistical information systems, couples very well with the possibility of handling the semantics of aggregation and generalisation that is possible in the context of an object-oriented data model . Moreover, it is in the context of the object-oriented approach that important reflections and formalisations of the structure and the semantic properties of the entities stored in statistical databases have been made.

The issue of capturing this functionality in the context of the temporal object oriented data models still needs investigation. Some type of metadata is needed, in the form of either inference rules or interpolation/extrapolation functions. For example, tuple time stamping using time intervals (or temporal elements) assumes that the values of the temporal attributes remain constant within an interval. If this is not the case, one must resort to explicit representation of each time point value, which may be either impossible or too expansive. Also if not all time points have stored values, some metadata information is needed in order to derive the data values for those time points. The way for synthesising and modelling this meta-information should be one of the several points of common interest for statisticians and database designers.

5. TEMPORAL QUERY LANGUAGES

The extension of the database approach to statistical applications has often been faced with the problem of the best semantics for statistical queries; a statistical query is, in general, more complex and it involves a large amount of data.
Relational algebra and relational calculus, in fact, do not formally incorporate the functions (like aggregation) that are necessary in statistical queries. These problems become obviously more serious when we want to manage the temporal dimension of data.
Semantics for a query language must have some properties that can be briefly summarised in this way: first of all it must be *declarative*, that is that it must assign a meaning to a query without referring to the way the query is evaluated; a query must be evaluated in a *closed form*, and the result of the query must be represented in the language of the database. This is trivial in the relational approach but for other kinds of databases, that contain for instance constraints this could become a non-trivial property. Again, a query language ought to be representation independent.
Some examples derived from the different proposals in the context of temporal query languages could be useful.
Tquel (Snodgrass, 1993) is a well-known temporal query language derived from Quel (Held, Stonebracker and Wong,1975). It supports a single temporal domain which is discrete, infinite and multi-level, (in the sense that it may handle days, hours, etc.) and two temporal dimensions: valid time and transaction time[4]. The data model of Tquel is a variant of the timestamp representation. In particular, a timestamp is an interval $<a,b>$ where a is the start and b is the end of the interval. Associating the interval $<a,b>$ with the fact $p(x)$ or with a tuple x in the relation corresponding to p, means that $p(x)$ holds for every $t, a <- t <- b$
The data model of Tquel is point based, not interval based, and intervals serve only as a representational device. The true values of facts are associated with points, not with intervals, and so it is impossible to represent a fact that is true in an interval but not in the different subintervals.
TSQL2 is a proposed extension to SQL2, and, like Tquel, is point based, and not interval based. Every fact has, in this sense, exactly one timestamp.
HSQL (Historical Query Language) (Sarda, 1990 (a), 1990(b),1993) refuses the cubic view of the database commonly used to capture the time dimension, and proposes a state-oriented view. The state of a database object is defined by the values of its attributes and its state prevails over an interval of time. The language is based on an interval stamping of tuples and it is based on an extended relational data model
Let us make a few final comments on the problem of incomplete or incompletely specified temporal information. This question is particularly delicate for statisticians that must explain/make explicit and handle missing data in their applications. For example, when the exact data of an event is unknown but it is known that it occurred before or after a certain date. In this case, at least from a theoretical point of view, the temporal languages would allow for modal queries to be posed, that result, in practice, in the possibility of distinguishing between facts that are *certain* and others that are *possible*.
To deal with incomplete temporal information, several authors have proposed to apply the classical framework to timestamp database and to model incomplete information in the context of the relational data model using marked nulls .
However the implementation of query languages for incomplete timestamp databases raises new problems because of the presence of timestamps formulas with nulls and the need for supporting modal queries. Some authors have studied, in fact, relational calculus, algebra and query languages for incomplete timestamp databases, but the complexity of query processing in such databases has been shown to be particularly serious.

6. CONCLUSIONS AND CORRELATED WORKS

The issue of a proper approach to temporal statistical information modelling, far from being only a computational or technical problem, is something deeply related to the level, the quality and the value of the information that can be derived from the collection of data.
Statisticians are users that are obliged to reflect with particular attention on this topic because they are involved in several ways. They are, of course, concerned as users, interested in the aspects of the independence of the data from the elaborative programs, but also as designers and users of statistical sources that, if correctly modelled, could increase their informative potentiality, and their efficiency. The point is, however, to succeed

[4]The *valid time* of a fact is the time when the fact is true in the modelled reality; a *transaction time* is the time when the fact is stored in the database.

in building information systems which allow for the use of expressive query structures to improve the quality of data, the integrity of the system, and the flexibility in organising the data in the most convenient way.
As we have seen, time introduces some further degrees of complexity in modelling. The problem of a correct management of time is nevertheless at the very core of the question of a correct organisation of statistical information, linked as it is to the problem of managing longitudinal data, survey data, information at different temporal granularity, detection of censored data, scientific databases, and metadata.
The challenge is now to suit methods and tools to the discovery of knowledge that will allow us to pass from information management to discovery. This problem is posed by the need to profit, even as statistical sources, from the huge amounts of data that have been stored in databases, in particular since the introduction of the relational model. While the data storage and handling mechanism have developed rapidly to cope with increasing volumes of data stored on computers, tools and methods for analysing them are far from satisfactory: databases are designed for purposes other than discovery and this fact poses a number of problems within the discovery process.
Finally, other problems are concerned with the need to incorporate data (saving their temporal context), from different databases and different sources within a single discovery process. For all these needs, the methodological and substantial contribution of the results achieved in the study of the mechanisms of knowledge and discovery and of artificial intelligence are pivotal.

Bibliography

Ariav G. A temporally oriented data model. *ACM Transactions on Database Systems*,11(4):499-527,December 1986

Atzeni P., De Antonellis V., *Relational Data Base Theory*,Bejamin Cummigs Publishing Company, 1993

Buzzigoli L., Martelli C., (a)Il rilascio di dati censuari: un confronto critico delle pratiche adottate dagli istituti nazionali di statistica, *Verso i Censimenti del 2000*, Udine 1999

Buzzigoli L., Martelli C (b).Le regioni come nodo di linkage di archivi amministrativi: limiti e potenzialità di un nuovo ruolo per la gestione di fonti statistiche, *Verso i Censimenti del 2000*, Udine 1999.

Chen P.P. The entity-relationship model: Toward a unified view of data.,*ACM Transactions on Database System* 1(1):9-36. March 1976

Codd E.F. A relational model for large shared dat banks. *Communications of the ACM* 13(6):377-387. 1970

Danmarks Statistik, *Personstatistik i Danmark: et Registerbaseret Statstiksystme*, 1994.

Elder G.H., Pavalko E.K., Clipp E.C., *Working with archival data- studying lives*, Sage University Press, 1993.

Gadia S.K., A homogeneous relational model and query languages for temporal databases. *ACM Transactions on Database Systems*,13(4):418-448, December 1988

Gadia S.K., Vaishnav J.H., A query language for a homogeneous temporal database. *Proceedings of the ACM Symposium on Principles of Database System*,:51-56, March 1985

Held G.D., Stonebracker M., Wong E., Ingres- A relational data base management system. In *Proceedings of the AFIPS National Computer Conference*, Vol. 44,Anaheim,CA, May 1975

Hinterberger H. Ed. Statistical and Scientific Database Management, *Proceedings of the VI Intern. Conference on Scientific and Statistic Database Management*, ETH Publ., Asona, Switzerland, June 6-10,1992

Inmon W.H., "Multidimensional Databases and Datawarehouse",*Data Management Review*, Jan. 1995

Kim W., Object-oriented Approach to managing statistical and Scientific Databases, in Z.Michalewicz(ed) *Statistical and Scientific Database Management*, Fifth International Conference,V SSDBM, Charlotte,N.C., USA, April 3-5,1990,Proceedings.

Kimball R., *The Data warehouse Toolkit*, Wiley & Sons,1996

Lorentzos N.A. , Johnson R.G., TRA: A model for a temporal relational algebra. *Proceedings of the conference of temporal aspects in Information Systems*, France, May 1987. AFCET.

Lorentzos N.A. , Johnson R.G., Extending relational algebra to manipulate temporal data. *Information Systems*,13(3), 1988

Lorentzos N.A., The interval extended relational model, in Tansel et *al Temporal Databases*, Benjamin/Cummings Publishing Company, 1993

Martelli C., Availability and limits of official data for demographic analysis based on biographies,in *Continuità e discontinuità nei processi demografici*, Arcavacata di Rende, 1995

Martelli C., Casini A., Strumenti per la raccolta di dati network., *Ingegnerizzazione del processo di produzione dei dati statistici* , Florence,April 1999

Martelli C.,(a) Basi di dati per lo studio delle biografie, in P. De Sandre, A.Pinnelli,A.Santini, *Nuzialità e fecondità in trasformazione. Percorsi e fattori di cambiamento*,Il Mulino,1999.

Martelli C.(b), Information Systems for a complex approach to demographic analysis, in Tabutin D., Gourbin C., Masuy-Stroobant G., Schoumaker B., *Théories, paradigmes et courant explicatifs en démographie*,L'Harmattan, 1999.

McKenzie E., Snodgrass R., An evaluation of relational algebras incorporating the time dimension in databases. *ACM Computing Surveys*,23(4):501-543,December 1991

Mc Clean S.I, Grossmann W. Froeschl K.A., Towards Metadata-Guided Distributed Statistical Data Processing, *NTTS'98- Internationaln Seminar on New Techiniques & Techonologies*,Eurostat, Sorrento**,** 1998

Michalewicz Z. Ed. *Statistical and Scientific Database Management*, Lectures Notes in Computer Science, N.420, Springer Verlag, 1990

Montanari A., Pernici B., Temporal reasoning, in Tansel et *al Temporal Databases*, Benjamin/Cummings Publishing Company, 1993

Navathe S.B., Ahmed R., TSQL-A language interface for history databases. In *Proceedings of the Conference on Temporal Aspects in Information Systems*:113-128, France 1987

Rafanelli M., Klensin J.C., Svenson P.,Eds, *Statistical and Scientific Database Management* , Lectures Notes in Computer Science, N.339, Springer Verlag,1989

Rose E., Segev A., TOODM- A temporal object oriented data model with temporal constraints. *Proceedings of the 10th International Conference on the Entity relationship approach*, October 1991.

Rose E., Segev A., *TO-Algebra- A temporal object oriented algebra*. University of California, Berkeley, January 1992.

Sarda N.L., Modeling of time and history data in database systems.*Proceedings CIPS Congress87* Winnipeg:15-20. CIPS,1987

Sarda N.L. (a),Algebra and query language for a historical data model. *The Computer Journal*,33(1):11-18,February 1990

Sarda N.L.,(b) Extension to SQL for historical databases. *IEEE Transactions Knowledge and Data Engeneering*, 2(2):220-230, July 1990

Sarda N.L, HSQL; A Historical Query Language, in Tansel et *al Temporal Databases*,Benjamin/cummings Publishing Company, 1993

Shoshani A., Wong H.K.T., Statistical and scientific database Issues, *IEEE Transactions on Software Engeneering*, 11(10):1040-10470, October 1985

Snodgrass R., An Overview of Tquel, in Tansel et *al Temporal Databases*,Benjamin/cummings Publishing Company, 1993

Tansel A.U., Clifford J., Gadia S., Jajodia S., Segev A., Snodgrass R., *Temporal Databases* Benjamin/Cummings Pub. Comp., 1993

Ullman J.D., *Principles of Database and Knowledge Base Systems, Vol. 1-2,* Potomac,Md.: Computer Science Press,1989

Vilain M.B., A system for reasoning about time. In *Proceedings of the American Association for Artificial Intelligence,* Pittsburgh, August 1982.