

Metadata Usage in Statistical Computing

Wilfried Grossmann

*Department of Statistics and Decision Support Systems,
University Vienna*

Email: wilfried.grossmann@univie.ac.at

Abstract

The statistical view towards data differs with respect to a number of aspects from the traditional view in Computer Science. Consequently, metadata for statistical data have to take into account these peculiarities in operational form. Based on a requirement analysis oriented towards statistical applications a generic metadata model is defined. Using the concept of facet classifications from the library sciences we define a model, which encompasses not only storage and retrieval aspects but also metadata usage in statistical processing. Within the framework of semistructured data modeling a process oriented data and metadata model for applications in statistical computing is sketched. Applicability of the model is shown in the context of weighting.

Keywords: Metadata, Semistructured Data, Statistical Data Models, Weighting

1. Introduction

Statistical computing as well as computational statistics start usually with rather simple data structures without taking into account many problems of applied statistics, which require besides the statistical methods a lot of data management and data manipulation as well as interpretation of results according to subject matter knowledge. In the area of official statistics this problem was recognized for a long time and lead to the development of the concept of statistical metadata (i.e. data informing about the data) (cf. Sundgren 1977). Due to the fact that these developments were more oriented towards computer science than statistics, recognition of these approaches by methodological oriented statisticians was rather limited.

It is the aim of this paper to show that also in the area of statistical computing metadata models can be utilized in order to support statistical analysis tasks. In section 2 we consider first of all a number of peculiarities of data occurring in the context of statistical investigations, which require specific features for statistical metadata modeling usually not considered in the traditional computer science approaches for metadata. Based on these considerations we outline a generic metadata model for statistical data. This model uses the idea of facet classifications, well known in the librarian community, and encompasses in that way besides storage and retrieval aspects also statistical processing aspects. In order to apply this model in the context of statistical computing we need a representation, which allows simultaneous processing of data and metadata. Section 3 discusses the main features of such a model for data as well as metadata, which is based on the theory of semistructured data. In section 4 we show the application of the approach in the context of weight calculations.

A major part of the results presented in this paper were obtained in connection with the METANET research project (IST-1999-29093), a thematic network within the statistical branch of the 5th research program of the European community.

2. Data and Metadata in Statistics

The term metadata is well established in the community of computer scientists and in the area of data archives and may be defined as any description of a data model (e.g. the schema) and the relationships between the components of the data model (cf. Jarke et al. (2000)). In order to obtain proper understanding of the requirements on statistical metadata we have to identify the peculiarities of data in statistics compared to the traditional view on data in computer science. Section 2.1 reviews some of these peculiarities. Based on these features we define a modification to the traditional data modeling scheme defined by the OMG standard (2000). In section 2.2 we outline a framework for metadata modeling in statistics which covers the identified statistical requirements.

2.1 Specific Features of Data in Statistics

Statistical data differ with respect to a number of aspects from traditional data. The most important differences seem to be the following ones:

a) Statistical Typology of Data

Statistics discerns in a natural way different types of data. The most important types are statistical populations, often represented as census, and sample or survey data collected for a specific purpose. Applied statistics refers usually to the latter but statistical methods can be applied also to entire, usually finite, populations, provided one takes into account the different nature of such data. Besides these two main types of data also other statistical data types occur in applications, in particular in context of data preparation prior to statistical computing in the narrow sense. Such data preparation steps require usually a mixture of statistical methodology and data management activities. A well-known example of advanced statistical methodology in this area is imputation for missing or incorrect values. But also computations, which are at the first glance of pure data management nature, for example aggregation of codes for quantitative variables, have to be done with a view towards its statistical implications and interpretations.

b) Production of Data

In ideal case production of statistical data is based on specific methods supporting interpretation of the results within the framework of statistical methodology. This framework allows statistical induction, i.e. generalization of the results from a sample to results for the underlying population. One of the challenges in new applications of statistical methods is how data obtained by non-statistical methods can be put into the framework of a statistical model. For example, Coppi (2002) defines a methodological format for data mining supporting the interpretation of statistical methods in this non-standard context.

c) Processing of Data

Statistics as a methodological discipline offers a number of tools for observational modeling, applicable to problems where either no, or only a partial analytical description of the system of interest is available. Exploratory data analysis and methods for condensing information about individuals to information about collectives may be in many cases a first step towards the development of more formal observational models. Such formal models are usually based on mathematical stochastic models, often combined in a clever way with other analytical models, for example general linear models. All these activities are different from the traditional data management activities considered in computer science.

d) Sources of Relationships Between Data

Data modeling in computer science defines relationships between data mainly by subject matter considerations, often modeled by entity relationship models for the data. Such relationships are not only important for correct interpretation of the data but can also imply application of specific statistical methods. Take as example the relationship between courses and students attending the courses, which may imply usage of a multilevel model. Besides such subject matter relationships we have to consider in statistics also the relationships defined by statistical methodology. Typical examples are correlation, relationships between sample and underlying population, or relationships between data and summary measures based many times on more complex operations than OLAP operations for databases.

e) Incompleteness of Statistical Data

As already remarked by Sato (1991) statistical databases violate the closed world assumption of traditional data bases, stating that all interested features about the object world are captured in the data base. In statistics we do not make such an assumption, hence we need two different conceptual schemes: one for description of reality and one for description of the actual data. In fact, statistical methodology offers a number of powerful tools for bridging the gap between these two schemes.

In order to take all these peculiarities into account it seems useful to modify the four layer scheme for data modeling proposed by the Object Management Group (see OMG (2000)), which defines these four layer by *data*, *definitions* (models of data), *methods for making definitions* (meta-models) and *methods that define methods* (meta-metamodels), in a way proposed originally in a data model for Bance d'Italia (Del Vecchio (1997)). This modified scheme defines a separate layer for statistical methodology and is shown schematically in Figure 1.

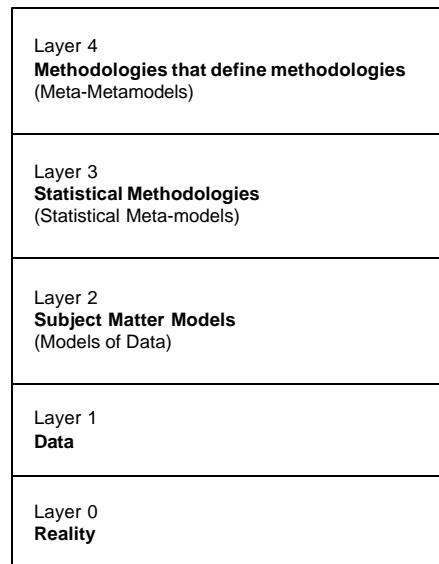


Figure 1. Schematic description of statistical data modeling hierarchies

The top (fourth) layer of this scheme encompasses general tools and methods and defines possible structural schemes, oriented towards computer science and general mathematical models. These tools and methods are at the disposal of the statistician in the realization of the models for statistical methodologies in the third layer, taking into

account the above mentioned peculiarities of statistical data. The second layer applies these statistical structures to subject matter problems and defines the subject matter data model for the data layer at the bottom (first layer). Additionally a zero layer representing reality is included.

2.2 Statistical Metadata

Statisticians, mainly in the area of official statistics were aware of all these problems and a number of research efforts have been made to define an appropriate framework for statistical metadata. In this context it is worth to mention that one of the first occurrences of the term metadata was by Bo Sundgren (1975) in the area of official statistics. In a recent proposal by Froeschl et al. (2003) a generic scheme for statistical metadata was defined within METANET, a thematic network project funded by the EU within the 5th IST research program of the European Union. The basic idea of this scheme is to keep data and description of data as close together as possible and to use metadata in an active way as a guidance and control structure for all statistical analysis activities. Such active use of metadata requires appropriate structuring of the data description. Following an idea well known in the community of librarians (Buchanan (1979)) we define a number of *metadata facets* described briefly in the following.

2.2.1 The Structure Facet

The structure facet defines the carriers of information in statistics, so called *categories*, resembling in a natural way the probabilistic setup for statistical analysis defined by probability space and random variables generating the event structure. Hence, the main categories are *statistical unit*, *statistical population* defined by the statistical units, *statistical variables* representing the process of measurement, *statistical values* defining the range (or co-domain) of the statistical variables, and *statistical datasets* defined by tuples of such statistical variables. The latter are of course the main object of interest in statistical analysis and may occur in different guise, for example as case level data (the well-known case by variates matrix), as summary level data (multidimensional tables), or vectors of observations (for example time series).

Besides these basic categories we need a number of additional categories describing the structure of statistical value sets: *grouping levels* defined by the aggregation of statistical values, *classifications* describing hierarchical schemes of grouping levels, *scales* capturing information about the measurement process and *measurement units* representing the meaning of quantitative variables. A further category called *statistical domain* is needed as organization principle for a statistical information system. Basically a statistical domain binds together the different categories occurring in connection with an investigation.

From data modeling point of view these categories can be interpreted as abstract classes. A concrete realization of a category occurring in connection with practical applications has to be represented in a twofold way: As category instance model (CI-model) and as category instance data (CI-data). The CI-data correspond to data occurring in the context of statistical computing and the CI-model describes the data. Obviously CI-data for the category statistical dataset are the most important type of data, but also other types of CI-data are well known: an administrative register may be seen as CI-data representing statistical units, a census file used in a sampling procedure may be seen as CI-data of a statistical population or a hierarchical classification of value sets used in a recoding step are the CI-data of a classification.

The CI-model is usually represented only partially in an extensional format as a ‘file description’ or a ‘codebook’, other parts of the description are many times only present in the mind of the statistician. Such an implicit consideration has often drawbacks with

respect to documentation of the analysis and implies many times also a lot of additional tedious activities, in particular in larger investigations done by different researchers. It is one of the main goals of this approach to overcome these obstacles by giving an explicit representation of all CI-data occurring in context of an analysis, together with the description of these CI-data by the corresponding CI-models. In ideal case such close connection between CI-data and CI-models allows the formulation of preconditions for statistical analyses in a more formalized way and supports in that sense the work of the statistician. Prerequisite for such an active use of CI-models in the analysis is representation in extensional format, i.e. as data. Such a representation of CI-models as data is called *statistical metadata*.

2.2.2 The View Facet

Formulation of CI-models in extensional format as data is based on a unified description principle captured in the so called view facet. It consists of four different views together with a structural description of the interconnections between the categories implied by the four views.

1. The Conceptual Category View

This view represents the subject matter definition of the category instance and builds in that sense the bridge to reality. Usually this view is represented by a verbal definition, in the most simple case an appropriate name or label for the object of interest. Besides this description we need in any way a temporal and geospatial specification stating time and location of validity for the definition.

Relationships between different category instances at the conceptual view resemble in some sense the traditional data modeling in form of ER-diagrams. Sundgren (1975) denoted the modeling of such relationships by the term infological approach.

2. The Statistical Methodological Category View

This view describes the objects of interest from a statistical point of view by a number of formal parameters. The most important parameters are *type parameters* characterizing the object of interest. For example in case of statistical datasets one type parameter specifies whether the dataset is based on case level data (often called microdata) or summary data (so called macrodata). In the former case information is given for each statistical unit in the underlying population whereas in the latter case information is given for classes of statistical units. Another type parameter for datasets is usually needed for determination of the temporal structure of the dataset, distinguishing between cross sectional and time series data. Note, that in general the values of such parameters may be combined in an arbitrary way, hence, the parameter concept is more flexible than the traditional inheritance structure in object oriented modeling.

Besides type specification we need also *role parameters* describing the actual role of the category in the context of a specific application. Contrary to the type parameters, which are usually fixed for a category instance during its lifetime captured by the stage facet (see subsection 2.2.3), role parameters may change during lifetime of the instance, even within one investigation. Well known examples for role parameters occur in connection with statistical variables: A variable may have in context of a statistical dataset the role of an identifier for statistical units (cases), the role of a cross classification variable identifying a class of cases, the role of a filter variable, the role of an explanatory variable in a model and so on.

Statistical relationships between the different categories are of utmost importance for establishing the statistical data model. For example, in case of statistical datasets the relationship to other categories may be described at a top level as shown in Figure 2. The relationships to statistical population and statistical unit are quite obvious and the structural relationships defined by the variables correspond to restrictions given usually

by the roles parameters of the involved statistical variables. Besides the relationships to the categories Figure 2 contains also some additional information about the production described in detail in section 2.2.3.

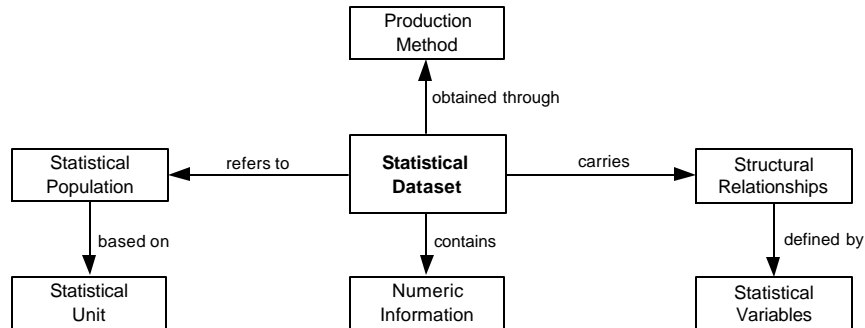


Figure 2. Relationships of a statistical dataset to other categories

3. The Data Management Category View

The *data management category view* is geared towards machine-supported manipulation, storage and retrieval of data. Main task of this view is management of CI-data in terms of files through properties often called *logistic metadata*. In general, the data management view concerns issues of how to represent, or encode, and manipulate entities and processes symbolically, especially in regard of storage and exchange, referring to data models and information structures as developed by computer science.

Depending on the specific category under consideration different elementary data structures are necessary:

- For statistical units a useful data structure would be a list with operations like insertion and deletion.
- For statistical populations and statistical value sets a set structure together with the standard set operations would be appropriate.
- In case of statistical datasets a number of different structures may occur: matrices in case of cross sectional case level data (the well-known case by variates matrix), multidimensional tables (cubes) in case of summary level data, or vectors in case of time series data.
- A special case of data structure is necessary for statistical domains as the basic organization principle of statistical systems. CI-data of a domain may be thought in the simplest case as catalogues of the different CI-models for the different CI-data used in the domain.

Due to the fact that statistical practice makes often explicit use of such elementary data structures, these structures are more important in statistical computing than complex data models for all the categories under consideration. This does by no means deprive the importance and usefulness of traditional data modeling, but within statistics these models can be applied only in a limited area in connection with a specific category view, for example an entity relationship model in order to describe the subject matter relationship between statistical units.

A detailed description of an XML representation of useful data structures, oriented towards R application, may be found in Meyer et al. (2002). The Data Documentation Initiative (DDI) (2001, see also <http://www.icpsr.umich.edu/DDI>) defines in its codebook a rather detailed description of such logistic attributes formulated in XML.

4. The Administration Category View

The administration category view addresses management and bookkeeping of all the structures. It is necessary for documentation of all kinds of activities in connection with definition of structures and schemas, insertion updating, and deletion of structures, as well as for search and retrieval activities. It has to take into account that production and storage of statistical data is often managed, or hosted, by public (often national) agencies or supranational (international) organizations with different subject matter orientation (e.g. economic data, social science data, biometric data, ...). This implies administrative structures exceeding by far the conventional horizons and functionalities of – more or less local – data (base) administration. Frequently, these structures also reflect *administrative* processes often implying that responsibility for data production and maintenance is spread among various administrative bodies, or agencies. Apparently, effective *statistical* usage of any of these data sources presupposes a fairly detailed knowledge of the administrative systems providing these data holdings. Moreover, legal aspects such as data privacy and data linkage prohibition rules have to be obeyed. While recent proposals for arranging data combination processes in the domain of data warehousing – as part of the so-called ETL-(extract-transform-load)-process – provide a range of technical solutions, administrative structures of data sources typically receive little attention.

A rather flexible attribute structure for documentation of administrative details has been worked out in context of the above mentioned DDI model, based on the Dublin Core standard for documentation of resources (see <http://dublincore.org/>). Although the intention of the DDI group is mainly documentation of statistical datasets it can be applied to all other categories with minor modifications.

2.2.3 The Stage Facet

The stage facet is responsible for support and documentation of all types processing of categories within the statistical information system. There exist a number of documentation templates for statistical datasets, which cover usually the entire processing chain. Two important approaches are the proposal by Rosen and Sundgren (1991) in the area of official statistics and the already mentioned DDI scheme for social science data archives. In both cases the documentation is more a passive metadata repository. In order to make active use of metadata in statistical computing one has to decompose these rather elaborated documentation templates into more operational building blocks. At the top level such decomposition can be defined by the four main stages: *definition and design*, *production*, *processing* and *dissemination and exchange*. A schematic representation of this decomposition is shown in Figure 3 and described in the following paragraphs.

1. Definition and Design

The definition and design stage defines the work plan for setting up CI-data in advance to the production of the data itself. It is based on the analysis of the object system and the intended mapping of the object system into statistical categories. Main result of the definition part is a so called *CI-blueprint* of the CI-models for the envisaged category instances, denoted by CM_i in Figure 3. These CI-models describe the intended CI-data according to the different category views described in 2.2.2, together with specification of the relationships to other category instances.

In case of statistical datasets the file description for data used by statistical analysis systems can be interpreted as sketch of a CI-blueprint referring mainly to the statistical and data management view.

Main result of the design part is an operational plan for the activities necessary for obtaining CI-data. Documentation of the planned activities is kept in a so called *CI-*

production-blueprint describing the activities according to the different category views. The methodological background for the design phase are sampling theory and planning of experiments, often considered only as a side branch in statistical computing. Consequently, there is often a gap in documentation of the CI-production blueprint in statistical software systems. Notable exemptions are specific tools for data capture and the already mentioned DDI standard, which keeps a quite complete documentation about the production steps.

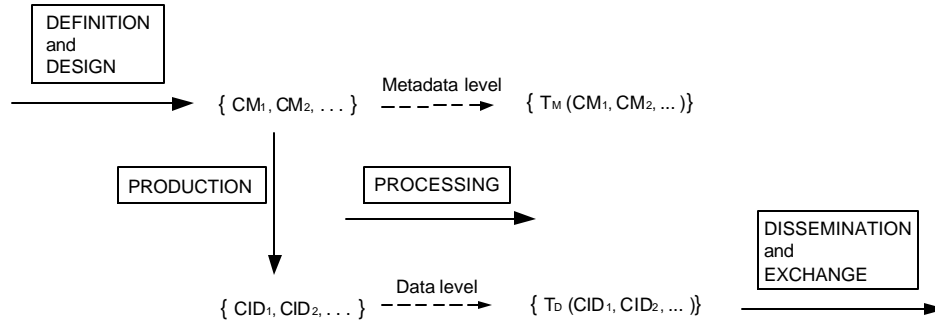


Figure 3. Main stages in the statistical processing chain

2. Production

The production stage establishes the CI-data for the category instances – denoted by CID_i in Figure 3 – according to the CI-production-blueprint and completes also the blueprint and the CI-models with respect to production dependent parameters. Such parameters may refer to sample sizes or non-responses in case of statistical datasets.

Obviously production of CI-data is mainly of interest in connection with statistical datasets and statistical populations. Theory of survey collection offers a number of instruments for computer assisted data production, for example software for questionnaire management. Modern data capture tools offer also a number of options for metadata management for the definition and design stage and use these metadata sometimes for active control of data capture. Furthermore, let us mention that a number of preprocessing steps, for example editing, are considered many times as part of the production stage.

Another example for documentation of the production step, mainly in the area of official statistics, occurs in context of classifications. International organizations put quite a lot of effort into documentation of the development of international standard classifications, usually in verbal form from the conceptual point of view.

3. Processing

The processing stage generates by a sequence of transformations new category instances out of already existing category instances. As shown in Figure 3 each transformation operates at the data as well as at the metadata level with strong interconnection between these two processing levels. Roughly speaking one can say that processing of the CI-models refer to preconditions and admissibility checks for processing at the data level, which corresponds more to statistical computing. Separation of these two levels implies often a gap in documentation and makes it difficult to reuse results and to give correct interpretation, in particular in case of larger projects.

Each transformation produces one or more output category instances from the already existing input category instances. A *planning phase* and an *execution phase* may be distinguished. The planning phase of a transformation defines an activity plan for the envisaged processing and as result of the planning phase one obtains one or more *CI-blueprints* and a *CI-processing-blueprint*.

The CI-blueprints, denoted by $T_M(CM_1, CM_2, \dots)$ in Figure 3, describe the CM-models for the output category instances according to the description of the statistical categories and the relationships between statistical categories as outlined in subsection 2.2.2.

The CI-processing-blueprint describes the envisaged processing activities for the CI-data – denoted by $T_D(CID_1, CID_2, \dots)$ in Figure 3 – with respect to the *conceptual, statistical processing, data management, and administrative processing view*:

- The *conceptual processing view* describes the processing method from a general point of view, in particular the intention of the method and the connection to subject-matter issues.
- The *statistical processing view* describes the transformation from a statistical methodological point of view. In particular, the following details have to be specified:
 - (i) The *input CIs* used in the transformation;
 - (ii) The *output CIs* produced by the transformation;
 - (iii) The *operators (statistical methods)* applied to the input CI-data;
 - (iv) The *operator parameters* necessary for detailed specification of the algorithm.
- The *data management processing view* is responsible for keeping additional results occurring besides the CI in connection with processing, for example process parameters.
- The *administrative processing view* informs about administrative details of processing.

In statistical computing main emphasis is on processing of statistical datasets using statistical analysis procedures. The planning phase corresponds in that case to the selection of the analysis procedure and log-files of procedure calls keep the essential part of the statistical processing view. Modern approaches in statistical software give also an explicit definition of the output objects, i.e. definition of the output category instances from statistical and data management point of view.

However, many times statistical computing is not limited to the manipulation of statistical datasets. A number of so called preprocessing stages in advance to statistical analysis corresponds to processing steps for CI-data of the categories statistical variables and statistical value sets, for example recoding of values or definition of new variables.

In the area of official statistics manipulation of statistical units and statistical populations plays also an important role. A well known example for statistical units is the production of so called analytical units out of already existing ones in order to obtain better comparable data. An example for processing of statistical populations is well known under the heading data combination.

4. Dissemination and Exchange

Dissemination and exchange are the main operations for obtaining information about category instances and processing activities of the category instances. All information required for dissemination is based on a specific view towards the category instances, the CI-production-blueprints and the CI-processing-blueprints.

Contrary to dissemination of CI-data, which is usually well defined, recommendations for the dissemination of CI-models is often rather vague. An attempt to describe the content of the CI-models for dissemination more precisely is given by the so called *function facet* of the metadata model. This function facet, not treated in detail in this

exposition, defines on the one hand the audience (users) of information, on the other hand content and format for dissemination. For example the UK data archive distinguishes between conceptual, contextual and cataloguing information. With respect to the format different specifications are well known in the area of official statistics, for example, GESMES (see <http://www.gesmes.org/>) for dissemination of statistical tables together with appropriate descriptive elements.

3. Metadata Structures for Statistical Computing

As already stated, statistical computing occurs mainly in the production and processing stage of the stage facet. In this section we consider only the application in the processing stage as described in section 2.2.3. In order to support simultaneously processing at the data as well as at the metadata level we need appropriate data structures. Definition of such data structures starts from the assumption that there exists a data and metadata store, organized in some data base, which is used as information source for definition of the computational plan from the conceptual point of view, i.e. the conceptual part of the processing blueprint. Based on this conceptual definition one can select the different CI-data necessary for computing, based on the data management view of the CI-model, together with that part of the CI-model relevant for computing. Usually this relevant information is kept in the statistical view of the involved CI-models. Hence, we restrict our considerations to the statistical point of view. Augmentation with the other views is rather obvious. After computation the results of computation have to be transferred back into the data and metadata store, which means definition of the data management point of view for the results as well as documentation of the results from a more conceptual and administrative point of view.

A rather flexible structure for keeping the information for computation at the data as well as at the metadata level for the categories was defined in Denk et al (2002) by a so called *composite structure*.

As generic syntactic building block of composites, a relational *container* structure is used defined by the following building blocks:

- (i) *buckets* holding all data extensions in connection with the categories;
- (ii) *bucket schemas* describing formal bucket structure;
- (iii) *directories* listing composite components.

Each composite comprises a *category directory*, a *container directory*, and, subject to its processing level, a different number of buckets and corresponding schemas.

Using *semi-structured* data modeling (see Abiteboul et al. (2000)), the formal set-up of composites is briefly sketched below. As usual, ‘*’ indicates an arbitrary number, ‘+’ at least a single occurrence of the preceding symbol; ‘?’ refers to optional occurrence. Prefixing with ‘&’ denotes references to sharable (typed) instances; ellipses indicate desirable extensions of composite definition.

Figure 4 shows the structure of a composite, resembling the description of the statistical methodological view in section 2.2.2. The context element specifies the role of the composite within the processing chain, defined by the stage facet (cf. 2.2.3). Within the ‘TypeParameters’ and ‘RoleParameters’ the notation *CTYPE* and *CROLE* is used as dummy notation for lists of admissible values depending on the specific category class. Obviously, also the composites referred to in the ‘StatisticalRelationships’ depend on the CategoryClass element. For example, in case of statistical datasets the relation refers to a population composite. The production step delivers usually ‘*source*’ composites, otherwise ‘*derived*’ composites generated by the referred transformation step are delivered. ‘CONTAINERDIRECTORY’ refers to a container component listing all containers the composite consists of.

```

type STATISTICALCOMPOSITE =
{
  (Label : string) ?,
  (Description : string) ?,
  Context : input | throughput | output,
  CategoryClass : unit | population | dataset | variable | grouping | ...,
  (TypeParameters : CTYPE)*,
  (RoleParameters : CROLE)*,
  DerivationOrigin : source | derived,
  GeneratedBy : &SOURCE | &TRANSFORMATIONSTEP,
  (Statistical Relationships : &COMPOSITE) ?,
  Components : CONTAINERDIRECTORY,
}

```

Figure 4. Type definition of a composite

In the description of the container directory in Figure 5 the referenced CATEGORYDIRECTORY will be of the type '*variables*', however in some cases also other types are possible. For example, in a composite representing statistical values the referenced directory may be of the type '*grouping*' defining correspondence tables obtained by recoding. '**R**' denotes a relational schema definition; parenthesized names are the relation's attribute with key attributes (left) being separated from non-key attributes (right) by a semicolon. The 'BucketRole' element identifies the role of the bucket within the composite. For example in case of a composite for statistical datasets the list '*BROLE*' may contain the values '*data*' for the data itself, '*sampling*' for the underlying sampling structure, '*weighting*' for possible weights of the datasets or '*results*' for the output data of a procedure.

The CATEGORYDIRECTORY itself consists of a container component listing all category instances in the composite and a reference to the composites it is used by.

```

type CONTAINERDIRECTORY =
{
  UsedCategories : &CATEGORYDIRECTORY,
  Contains : R( BucketRole : BROLE ;
                Schema : &SCHEMA,
                Bucket : &BUCKET )
}

type CATEGORYDIRECTORY =
{
  Contains : R( ID : category_id;
                TypeParameter : CTYPE ,
                Role : CROLE ,
                CorrespondsTo : &CONTEXTATTRIBUTE ,
                ... ),
  (UsedBy : &COMPOSITE) +
}

```

Figure 5: Type definitions of container directories and category directories

Figure 6 shows the type definition for buckets and for bucket schemas. Main part of the bucket schema is the relation listing bucket attributes and references to the

composites it is used by. In addition, the role of the bucket and the format of the bucket, i.e. the used data structure, it describes, are stated.

In the bucket type definition '*BUCKETDATA*' is to be substituted for the actual data relation as determined by bucket schema '&SCHEMA'.

```

type SCHEMA =
{
  BucketFormat : BFORMAT,
  BucketRole : BROLE,
  Contains : R ( ID : bucket_id;
                CorrespondsTo : category_id,
                ... ),
  (UsedBy : &COMPOSITE) +
}

type BUCKET =
{
  Schema : &SCHEMA,
  (UsedBy : &COMPOSITE) +,
  Contains : BUCKETDATA
}

```

Figure 6. Type definitions of buckets and schemas

Composites provide the operand structure of statistical processing. Processing itself is based on a number of transformation steps. Figure 7 shows the type definition of a transformation step resembling the statistical processing view of the CI-processing-blueprint

```

type TRANSFORMATIONSTEP =
{
  (Label : string) ?,
  (Description : string) ?,
  Applies : &OPERATOR,
  Uses : PARAMETERS,
  (Input : &COMPOSITE) +,
  Output : &COMPOSITE,
  (Transforms : &CONTEXTATTRIBUTE) *,
  (Generates : &CONTEXTATTRIBUTE) *
}

```

Figure 7. Type definition of transformation steps

Essentially, for each operator, (i) a structure holding *parameters* of the operator, (ii) *preconditions* concerning these parameters as well as (iii) *post-conditions* on the structure of the output composite, must be defined. For the description of parameters, again, semi-structured data typing can be used, allowing a formal definition of these conditions as path constraints in the data model. Each individual transformation step, i.e. each call to an operator, creates a log-entry containing references to the applied operator, to input and output composite(s) as well as to further parameters and, incidentally, transformed or generated categories of the shared transformation context. This way, a transparent record of transformation sequences is established.

4. Application Example

Application of the model in statistical computing is usually done in three steps. First of all one has to make a requirement analysis defining the scope of the statistical problems. Based on this requirement analysis one has to specify the necessary data structures in detail. In particular one has to specify all categories modeled as composites, together with the admissible values for the parameter lists. In a third step one has to formulate the transformation steps together with the checks for the preconditions for the operators. In order to show the applicability of this general plan we use as example computation of weights (see Grossmann and Ofner (2002)). An implementation of the model could be done in different environments. In order to be fully compatible with some standard in statistical computing a prototype was implemented in the SAS environment, i.e. all structures described in section 3 are realized as SAS datasets. Such an implementation is convenient for performing statistical computing with the data but rather cumbersome with respect to data administration inside the structure. Details may be found in the thesis of Ofner (2001).

4.1 Requirement Analysis

From a statistical point of view a number of approaches towards weighting are possible. The first one is the design-based approach assuming known sampling probabilities for the observation units. Based on the sampling probabilities p_{is} for the observed units one can define the weights proportional to $1/p_{is}$. Such weights are often called *base weights* valid for all variables in the survey dataset and the corresponding estimator is known as Horvitz-Thompson estimator.

The second approach closer to traditional statistical modeling is the model-based approach. It assumes that values of *auxiliary variables* $X = (X_1, X_2, \dots, X_p)$ are available for all population units prior to sampling. Based on the auxiliary information one defines a linear model for the target variable Y of the following form: $E[Y] = Xb$, $Var(Y) = V$. Using the observations y_s of the target variable one can estimate the parameter vector b from the sample and define an optimal estimator for the quantity of interest (cf. Valliant et al. (2000)). Usually these weights are specific for each target variable Y .

A third alternative is the model-assisted approach, which tries to bridge the two alternatives by using information from the sampling design for calculation of base weights and auxiliary variables for modification of base weights. These new weights are calibrated in such sense that the weighted sums of population totals for the auxiliary variables reproduce the known population totals. Depending on the type of the auxiliary variables a number of computational procedures are available. Well known examples are generalized regression estimates (GREG), based on a universal regression model for all target variables, or calibration weights, defined by minimizing a predefined distance between base weights and the new calibrated weights (Deville and Särndal (1992)). In case of qualitative auxiliaries a rather simple calculation method known as raking can be used (cf. Kalton et al. (1998) for an introductory survey). Usually these weights are considered as universal for the whole data set.

From the exposition it is obvious that calculation of weights presupposes the combination of data from a number of different sources: besides survey data we need also data for the sampling design and data for auxiliary variables from a population database. These data sources are often not fully compatible with respect to their structure, for example case by variates survey data may be used together with a sampling plan defined

in a table of selection probabilities for strata, and auxiliary variables represented in tables of marginal counts for the population. Consequently, calculation of weights needs a number of preprocessing steps for aligning the data according to the requirements of the statistical algorithms.

4.2 Data Structures

Corresponding to the requirement analysis we discern composites for statistical data and statistical populations. In both cases the categories referenced in the container directory are statistical variables. Inside the composite for statistical datasets we have buckets of four different bucket roles: a *data bucket* for survey data, a *sampling bucket* for sampling information, a *weight bucket* for the weights available for the dataset and a *method bucket* for the results of the calculations applied to the data. For the type parameter element in the category directory of the statistical variables we use a conventional specification like *quantitative*, *qualitative* or *string* variables. The role of the variables within the composite depend on the bucket role. Typical roles in a data bucket may be *identifying* variable (key inside the bucket), *auxiliary* variable or *observation* variable. In the sampling bucket possible roles are *selection probability* or *stratum count*. The roles of variables in a weight bucket correspond to the weighting method and inside a method bucket the roles are defined according to the applied transformation. A typical role in the context of weighting may be variance of the estimate for the population total.

Figure 8 shows as example the relation in the 'Contains' element of the bucket schema for the data bucket. As an example for the ellipses in the general definition of Figure 6 we have used the relational attribute 'GroupingLevel' defining the selected range of the statistical variables. The grouping levels support a number of predefined recoding and transformation operations for variables (see Papageorgiou et al. (2001) for details in a similar model). With respect to the 'BucketFormat' element we distinguish between *case* and *summary* indicating whether the dataset itself is a case by variates matrix or a summary table.

ID	CorrespondsTo	GroupingLevel
id (key)	V1	-
sex	V2	0
education	V3	1
income	V4	1

Figure 8. Example of bucket schema for data

Figure 9 shows an example for a sampling bucket with variable roles indicated in the header of the columns: The qualitative variable V5 defines the strata, V6 the number of observations in each stratum and V7 represents the sampling plan.

V5 (stratum id)	V6 (stratum count)	V7 (selection probability)
Stratum 1	880	0,25
Stratum 2	144	0,30

Figure 9. Example of a bucket for sampling

The population composite is structured similar to the data composite: the *data bucket* is reserved for population data available either as a case level data (BucketFormat *case*) or as marginal counts (BucketFormat *summary*). The *structure bucket* informs about the population structure, usually defined by stratification.

4.3 Computational Procedures

Following the general setup of Figure 7 description of the transformation encompasses besides the input and output composites the definition of an operator together with specification of operator parameters, a variable list to which the operator applies and a variable list which is generated by the operator. Based on these specifications the transformation is executed by the following main processing activities:

- (i) Feasibility check for the transformation;
- (ii) Definition of a detailed computation plan;
- (iii) Statistical computation according to the plan;
- (iv) Generation of the output composite and documentation.

The feasibility check infers whether the required computation is possible using information contained in the bucket schemas and the description of container attributes, in particular the roles of the variables inside the composite. Checking presupposes a careful analysis of the envisaged operations from a statistical point of view. In case of weighting we have to analyze the data requirements for different types of weighting. Let us consider two examples:

- (i) Calculation of base weights is feasible, provided that the variables describing the sampling design in the sampling schema are available also in the data container and the grouping levels are in both containers compatible. For instance in the example given by Figures 8 and 9 base weights are not feasible, because variable V5 is not part of the bucket schema for the data.
- (ii) Checks for calibration weights have to take into account the type of the auxiliary variables. Consider for instance the case of two categorial auxiliaries and a desired adjustment according to the marginals with respect of these two variables (defined in the parameter specification of the transformation). Calibration is only feasible if the two auxiliaries are used in the data container in compatible form and calculation of base weights can be done in advance.

From computer science point of view such checks can be treated formally using the parameter specifications of the composites. In the SAS prototype these checks were realized as SAS-macros.

Based on the results of the feasibility check one can define a detailed computation plan by specifying a sequence of conventional statistical algorithms. For example, in the above sketched case of calibration weighting the plan encompasses following steps:

- (i) Calculation of base weights using data contained in the sampling container;
- (ii) Calculation of calibration factors using data from the population composite;
- (iii) Determination of weights in a form compatible to the structure of the survey data in the statistical composite.

References

- Abiteboul, S., Buneman, P. and Suciu, D. (2000), "Data on the Web / From Relations to Semistructured Data and XML", Morgan Kaufmann Publishers, San Francisco.
- Buchanan, B. (1979), "Theory of Library Classification", Clive Bingley, London.

- Coppi ; R. (2002), “A theoretical framework for Data Mining: the Informational Paradigm”, *Computational Statistics & Data Analysis* 38, 501 – 515.
- Data Documentation Initiative (DDI) (2001), “Codebook Document Type Definition (DTD)”, <http://www.icpsr.umich.edu/DDI/CODEBOOK/>.
- Del Vecchio, V. (1997), “La rappresentazione dei dati e dei concetti statistici – Tematiche haziendali”, Internal Report, Banca d’Italia (see also Froeschl, Grossmann, DelVecchio 2003).
- Denk, M., Froeschl, K.A., Grossmann, W. (2002), “Statistical Composites: A Transformation-bound Representation of Statistical Data”, in *Proceedings 14th Conf. Scientific and Statistical Database Management*, ed. J. Kennedy, ACM SIGMOD, Los Alamitos, pp. 219 – 226.
- Deville, J.C., Särndal, C.E. (1992), “Calibration estimators in survey sampling”, *J. American Statistical Association*, 87, 376-382.
- Froeschl, K. A., Grossmann, W., Del Vecchio, V. (2003), “The Concept of Statistical Metadata”. Deliverable 5 METANET (EPROS Project IST- 1999-29093), University of Edinburgh.
- Grossmann, W. Ofner, P. (2002), “A Self Documenting Programming Environment for Weighting”, in *Proceedings in Computational Statistics (COMPSTAT 2002)*, eds. W. Härdle, B. Rönz, Physica, Berlin, pp 129 – 134.
- Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P. (2000), “Fundamentals of Data Warehouses”, Springer, Berlin.
- Kalton, G. Flores-Cervantes I. (1998), “Weighting Methods”, in: *New Methods in Survey Research* (A. Westlake et al. eds.), Association for Survey Computing, Chesham Bucks, UK, pp. 77-93.
- Meyer, D., Leisch, F., Hothorn, T., Hornik, K. (2002), “StatDataML: An XML Format for Statistical Data”, in *Proceedings in Computational Statistics (COMPSTAT 2002)*, eds. W. Härdle, B. Rönz, Physica, Berlin, pp 545 – 550.
- Ofner, P. (2002), “Embedding of Weighting Algorithms into Metadata Structures”, Thesis, Univesity Vienna.
- OMG (2000), “Common Warehouse Metamodel Specification”, OMG Document, <http://www.omg.org/news/releases/pr2000.htm>
- Papageorgiou, H. Pentaris, F. Theodorou, E. Vardaki, M. Petrakos, M. (2001), “A statistical metadata model for simultaneous manipulation of both data and metadata”, *J. of Intelligent Information Systems* 17, 169 – 192.
- Rosen, B. Sundgren, Bo (1991), “Documentation for reuse of microdata from surveys carried out by Statistics Sweden (SCBDOK)”, *Statistics Sweden*.
- Sato, H. (1991), “Statistical Data Models: From a Statistical Table to a Conceptual Approach”, in: *Statistical and Scientific Databases*, ed. Michalewicz Z., Ellis Horwood, Chichester, pp. 167–200.
- Sundgren, Bo (1975), “Theory of Databases”, Mason/Charter, New York.
- Sundgren, Bo (1977), “Meta-Information in Statistical Agencies”, *Conference of European Statisticians, ISIS’77 Seminar*.
- Valliant, R., Dorfmann, A.H., Royall, R.M. (2000). “Finite Population Sampling and Inference”, *Wiley Series in Probability and Statistics*, New York.