

Intelligent Support for Solving Classification Differences in Statistical Information Integration

C.M. Jonker¹ and D. Verwaart²

¹Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam
jonker@cs.vu.nl

²Agricultural Economics Research Institute LEI, Burg. Patijnlaan 19, 2585 BE den Haag
d.verwaart@lei.wag-ur.nl

Abstract. Integration of heterogeneous statistics is essential for political decision making on all levels. Like in intelligent information integration in general, the problem is to combine information from different autonomous sources, using different ontologies. However, in statistical information integration specific problems arise. This paper is focussed on the problem of differences in classification between sources and goal statistics. Comparison with existing information integration techniques leads to the conclusion that existing techniques can only be used if individual data underlying the statistics is accessible. This requirement is usually not met, due to protection of privacy and commercial interests. In this paper a formal approach and software tools are presented to support statistical information integration, based on a generic ontology for descriptive statistics, and heuristics that work independent of the domain of application. The heuristics were acquired from economic experts working in the field of European Common Fisheries Policy.

1 Introduction

Statistics are indispensable for political decision making. Economic, demographic and environmental statistics are used for monitoring social and physical processes and for measuring policy effectiveness. National governments usually have organised statistics services in order to fulfil their demand for decision support. In supranational organisations like the European Commission homogeneous statistics are often not available. Organizing supranational statistics is a time-consuming and precarious task. In many cases political decisions have to be made and processes have to be monitored long before homogeneous statistics can be available. Then heterogeneous statistics from a variety of independent sources must be integrated.

European Common Fisheries Policy (CFP) [1] is an example of a field where insufficient homogeneous statistics are available. Annually an economic report is prepared by a group of experts from the involved countries [2]. The statistics in this report integrate a broad variety of national and regional statistics. This task is performed in two annual workshops. In spring, the experts meet in order to agree on the contents of the report and to plan the data collection. Back home they collect the best available data. In the autumn workshop the data is integrated. Providing

automated support for the integration process would make more time available for economic analysis in the workshop, and improve the quality of the statistics because under pressure of time the experts sometimes make errors.

Over the past ten years much has been achieved in the field of intelligent information integration. Examples of approaches that support integration of quantitative data are SIMS [3], COIN [4], HERMES [5] and InfoSleuth [6]. None of these systems, however, supports integration of statistics by explicit mechanisms for mapping aggregated data from a particular classification to a target classification, as described in section 3. Most current research is focused on extending integration techniques from structured databases to semi-structured data on the web, as can be illustrated by Ariadne [7]. In integration of statistics, all general problems known from other areas of information integration occur, such as ontological and notational differences and differences in units of measurement and typology. In addition, specific problems in the integration of statistics are:

- differences in population, e.g., differences in threshold for inclusion of objects. For example, does a boat with engine power less than 20 hp count as a fishing vessel?
- differences in reported statistics, e.g., sum versus average.
- differences in classification, e.g. age classes bounded by 20, 35, 50 and 65 years vs. 15, 35 and 55 years; or length vs. gross register tonnage as vessel size indicator.

This paper concentrates on the fundamental problem of classification differences.

The specific problems of integration of statistics occur only if the underlying individual data sets are inaccessible. If all individual data would be accessible, available integration support systems like SIMS [3] could be applied and the integrated results could be aggregated to the desired level of specification. In many cases individual data is inaccessible for reasons of privacy or commercial protection. Therefore, dedicated methods have to be developed for integrating statistics.

One of the problems underlying statistics is that semantics of statistics are not included explicitly in the statistics itself, nor are they universally defined. The specific semantics of a statistic are often obtainable, but, in general, this is not a trivial process. Integration of several heterogeneous statistics requires the understanding of the semantics of each of the statistics, heuristic knowledge about the domain of application, and a general understanding of the discipline of statistics. Therefore, techniques for automated support for statistical information integration require the explicit use of heuristic knowledge and cannot be seen as a mere statistical problem and the implementation of a statistical technique. This heuristic knowledge must be acquired from human experts in the field.

In this paper a generic model is presented for resolving classification differences, applying domain heuristics. The model contains a generic ontology of descriptive statistics, and an explicit model of a specific method, the “weight matrix method”. The approach exploits in a generic manner domain specific heuristic knowledge about relationships between statistical variables. As a result the approach and its software can be used on any domain, given arbitrary (domain specific) sources and heuristic knowledge. The approach has been implemented in a prototype.

In section 2 the problem addressed in this paper is elaborated. In section 3 the approach to solve classification differences is explained and motivated by the experience of human experts. The model is presented in section 4. Section 5 gives an example of the application of the model. The results are discussed in section 6.

2 STATISTICAL HETEROGENEITY

In this section the concepts of homogeneity and heterogeneity are defined with respect to statistics, and classification difference is introduced as one of the causes of statistical heterogeneity.

A statistic is a non-empty set of values of one or more statistical variables, specified over zero, one or more dimensions. A statistical variable is a variable that describes an aggregated property of a population of objects or of one or more subclasses of a population. A subclass is a population subset, defined by the value of one or more variables. The classifying variables are referred to as dimensions of the statistic. An example of a statistic is a table describing total catch of fish (the statistical variable) per species per week (the dimensions) for the North-sea fishing fleet (the population).

A set of statistics is semantically homogenous, if

1. all objects represented in the population are subject to the same inclusion criteria,
2. all statistical variables that describe the same property have equivalent definitions,
3. all subclasses either overlap completely or do not overlap at all, and
4. all variables classifying for equivalent dimensions have equivalent definitions.

A set of statistics is called semantically heterogeneous if at least one condition for homogeneity is not satisfied. Following this definition, statistics are semantically heterogeneous if they differ in what they *intend* to describe. Differences in data collection process and data processing are excluded from the definition of semantic heterogeneity, although they may cause data inconsistency. These differences are excluded from the definition because they apply to the way statistics are produced, just like other causes of inconsistency (omissions, errors and fraud).

In order to integrate statistics that are heterogeneous only with respect to their dimensions, they have to be mapped from their original source classification to a common target classification. Two kinds of mappings are used in statistical information integration: individual and aggregated mappings. In the process of mapping data about individual objects, object identity is preserved: only attribute values are mapped to values of other attributes. For aggregated statistics object identity is not preserved in the mapping. The following example illustrates these issues. Assume that statistic S distinguishes between large and small vessels at class boundary 40 m, and statistic T also distinguishes between large and small vessels, but at class boundary 50 m. A 45 m vessel is classified “large” in S and “small” in T, but it remains the same vessel: it has the same identity in both S and T. The class of large vessels in S, however, is different from the class of large vessels in T, and comparing them requires information about some vessels that are called small in T.

3 RESOLVING CLASSIFICATION DIFFERENCES

Classification differences can be solved by finding an appropriate matrix W that relates the values of a statistical variable according to different classification:

$$WC=D \tag{1}$$

where C and D are $n \times 1$ and $m \times 1$ matrices respectively, representing the values according to the respective classifications, and W is an $m \times n$ matrix representing the relationship. If the statistical variable is the total value (sum) of a population variable for each class, matrix W is a weight matrix representing the distribution of C over the classification of D . In the remainder of this section the approach human experts take and a general recipe for the weight matrix method are described.

An interesting case is integration of data about Belgian fisheries. Two departments of the Belgian ministry of agriculture and fisheries published statistics on the aspects of fisheries they are responsible for: landings (amount of fish brought ashore) [9] and financial results [10] respectively. Both aspects are included and compared in the annual economic reports for the CFP. However, the two Belgian reports use different vessel classifications. The approach taken by human experts (fisheries economists) is to redistribute the total value of landings per class from [9] over the classification of [10]. To create a weight matrix, the experts used known fishing effort data, specified for the cross product of both classifications. So, fishing effort (in kWdays) is used as a proxy variable for value of landings (the term proxy is used in econometrics for a stand-in variable that is approximately proportional to an unobservable variable [11]).

The recipe for the weight matrix method, as given by the experts, is as follows:

1. Primary source determination:
 - Find the sources that contain the requested statistical variable (or a variable that can be transformed to it) for the requested population.
 - If a source exists of which the classification matches the requested classification, homogeneous data is available that can serve as goal table.
 - Else, select the most reliable source as primary source.
2. Weight source determination: if the request is not yet satisfied, find a source containing the source classification variable for the primary source, the goal classification variable, and a variable that can be used as proxy variable for the requested statistic.
3. Construct the weight matrix.
4. Multiply primary source with weight matrix to compute the goal statistic.

This recipe is applicable for transforming sum data. In order to transform other statistics, e.g. average data, the experts transform to sum data before applying the recipe. The latter transformation process is not in the scope of this paper. Steps 2 and 3 are the most complex. Step 2 requires proportionality models describing the trust that experts have in usability of variables as proxy for other variables. Step 3 entails aggregation of the proxy variable from individual data, according to the cross product of both classifications. These steps are further explained in the next section.

4 STATISTICAL SUPPORT MODEL

In this section the process model for statistical support is introduced. The compositional development method DESIRE [12] was used to specify and implement the model. The model basically follows the steps of the weight matrix method as introduced in the previous section (see Figure 1). The model includes a statistical ontology that uses generic terms and relations common to statistics. First the ontology

is introduced, and then the components of the process model are explained. The ontology elements occurring in this paper are:

```

relations:
goal_population: SN;
goal_variable: VAR;
goal_classification_variable: VAR;
source_classification_variable: VAR * C_DESCRIPTION;
goal_classification: VAR * C_DESCRIPTION;
gc_description: C_DESCRIPTION;
goal_type: TYPE;
goal_table: SN;
contains: SN * VAR;
source_aggregation_level: SN * AGRR_LEVEL;
describes_population: SN * POPULATION;
proportion_model: P_MODEL * TRUST
possible_sc_var_wrt: CLASSIFICATION * SN;
candidate_weight_source: SN;
possible_weight_tuple: VAR * SN * TRUST;
terms:
i_class(...) /* interval class */
class_var(...) /* population variable used in classification */

```

where SN is the sort of source names, VAR is the sort of statistical variables, population variables, and classifications. The sort TYPE describes the type of statistic (e.g., total, mean). Classification descriptions (e.g., list of intervals) belong to sort C_DESCRIPTION. Aggregation levels (individual, aggregated) are indicated by sort AGGR_LEVEL. POPULATION is the sort of individuals described by a statistical source. The sort of P_MODEL contains descriptions of proportionality models. The sort CLASSIFICATION is the sort consisting of objects built up out of a variable name and a C_DESCRIPTION. The sort TRUST indicates the confidence experts have in a model.

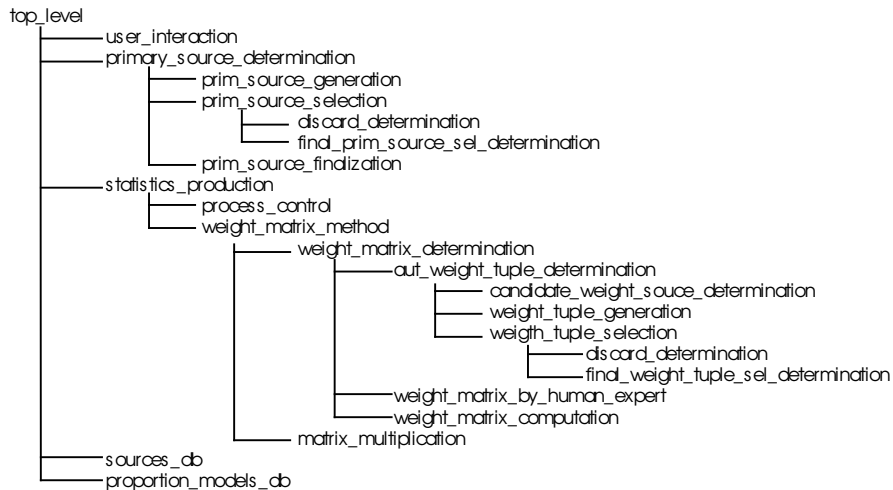


Fig. 1. Process Composition

Component `user_interaction` contains the interfaces necessary to obtain requests and to present results. It analyses the request in terms of goal population, goal variable, and goal classification. Components `sources_db`, and `proportion_models_db` refer to databases containing all available source descriptions (and links to XML documents containing their data) and proportionality models.

Component `primary_source_determination` corresponds to step 1 of the weight matrix method. Its task is to determine the primary source, containing the goal variable. It consists of three processes: generation (`prim_source_generation`), selection of the most reliable candidate (`prim_source_selection`), and collecting associated, necessary facts of the selected source (`prim_source_finalization`). Only the knowledge used in the generation component is described here.

```

if      goal_variable(G: P_VAR)
and    goal_type(T: TYPE)
and    goal_population(P: POPULATION)
and    contains(S: SN, stat_var(T: TYPE, G: P_VAR))
and    contains(S: SN, SC: CLASSIFICATION)
and    describes_population(S: SN, P: POPULATION)
then   possible_sc_var_wrt (SC: CLASSIFICATION, S: SN);

/* optimal case: SN satisfies the request */
if      goal_classification_variable(X : P_VAR)
and    gc_description(CD: C_DESCRIPTION)
and    possible_sc_var_wrt
      (class_var(X: P_VAR, CD: C_DESCRIPTION), S: SN)
then   goal_table(S: SN);

```

The component `statistics_production` is set up to be extendable with more methods for production of statistics, e.g. transformations between sum and average. For the same reason the component `process_control` is included. In this paper, the focus is on component `weight_matrix_method`, which consists of `weight_matrix_determination` (steps 2 and 3 of Section 3.2) and `matrix_multiplication` (step 4).

The component `weight_matrix_determination` consists of `weight_matrix_computation`, `aut_weight_tuple_determination`, and `wm_by_human_expert`. If the automated construction of the weight matrix fails, a human expert can enter a matrix directly. The automated process starts with `aut_weight_tuple_determination`. That process is composed again (see Figure 1). The first step is to select candidate weight sources.

```

if      source_classification_variable(X: P_VAR)
and    goal_classification_variable(Y: P_VAR)
and    goal_population(P: POPULATION)
and    contains(S: SN, X: P_VAR)
and    contains(S: SN, Y: P_VAR)
and    describes_population(S: SN, P: POPULATION)
then   candidate_weight_source(S: SN);

```

The next step is to select candidate proportionality models in order to obtain possible weight tuples (triples of proxy variable, weight source and trust):

```

if      goal_variable(G: P_VAR)
and    proportion_model(input_output_vars
      (I: P_VAR, G: P_VAR), P: MODEL_PRECISION)
and    candidate_weight_source(S: SN)

```

```

and contains(S: SN, I: P_VAR)
then possible_weight_tuple(I: P_VAR, S: SN, P: TRUST);

```

From the triples, one is selected with the highest trust (a quality indication for the proportionality, provided by experts). If this succeeds, a precursor of the weight matrix is computed using (XML) data from the weight source. The elements of the precursor are computed as sums of weight values by adding up the input values from the weight source for which the classification variable values match both the corresponding goal class and the corresponding source class. The matching criterion for interval classes is $\text{lower_bound} \leq \text{classification_variable_value} < \text{upper_bound}$. For other types of classifications the criterion is semantic equivalence of strings.

The precursor $m \times n$ matrix $[w'_{ij}]$ is normalized by dividing each element by the sum of all elements matching the same goal class, in order to produce a weight $m \times n$ matrix $W [w_{ij}]$:

$$w_{ij} = w'_{ij} / \sum_{i=1}^m w'_{ij} \quad (2)$$

The next step in the process is extraction of (XML) data from the primary source to create the $n \times 1$ input data matrix C obtained by taking the relevant vector from the primary source with respect to the source classification variable. The last step is the multiplication WC to obtain the goal table.

The transformation method described in this section transforms sum data for one variable of one population. To create an integrated statistic, it has to be executed for every combination of goal population and goal variable in the request. Furthermore, the transformation of average to sum data (and reverse) may be required. The next section gives a detailed example for a simple transformation.

5 EXAMPLE TRANSFORMATION PROCESS

Length is used as an indicator for ship size in German fisheries statistics. In Danish statistics, gross register tonnage is the common indicator for ship size. Suppose the following goal table is requested in order to compare German catch with Danish data:

Total catch of the German fleet in 2000, by gross register tonnage (GRT) class

GRT class (tons)	total catch (tons)
0-40	...
40-400	...
400 or more	...

Internally, this request is formulated as follows:

```

goal_population('German fleet 2000');
goal_variable(stat_var(total, 'catch'));
goal_classification('GRT_class',
  [i_class('0-40', 0.0, 40.0),
   i_class('40-400', 40.0, 400.0),

```

```
i_class('400 or more', 400, ANY)];
```

No source is available that can provide this data. However, a primary source (the German “Fangstatistik”) exists for total catch by ship length:

Total catch of the German fleet in 2000, by length class	
Length over all (m)	total catch (tons)
0-10	6268
10-20	30223
20-50	54131
50 or more	116594

The primary source description is as follows:

```
contains('total catch Germany 2000', class_var('LOA',
  [i_class("0-10", 0.0, 10.0),
   i_class("10-20", 10.0, 20.0),
   i_class("20-50", 20.0, 50.0),
   i_class("50 or more", 50.0, ANY)]));
contains('total catch Germany 2000', stat_var(total, 'catch'));
source_aggregation_level
  ('total catch Germany 2000', aggregated);
describes_population
  ('total catch Germany 2000', 'German fleet 2000');
```

The following weight matrix is required in order to compute the goal table.

GRT	Length over all			
	0-10	10-20	20-50	50+
0-40
40-400
400+

In the databases, the German fleet register of 1998 is available, giving length over all and gross register tonnage for each ship. This satisfies the constraints for a candidate weight source: it contains source classification variable and goal classification variable, and describes a population similar to the goal population. The relevant description elements are:

```
contains('Fleet register Germany 1998', 'LOA');
contains('Fleet register Germany 1998', 'GRT');
source_aggregation_level
  ('Fleet register Germany 1998', individual);
describes_population
  ('Fleet register Germany 1998', 'German fleet 2000');
```

The source ‘Fleet register Germany 1998’ is selected as a candidate weight source. Fisheries expert knowledge about proportionality between catch and engine power is formulated as follows:

```
proportion_model
  (input_output_vars ('engine power', 'catch'), 0.7);
```

The factor 0.7 indicates the confidence that experts have in using engine power as a proxy for catch; 0 would indicate complete distrust, 1 complete trust. This model matches a variable occurring in a candidate weight source with the goal variable. Combined with 'Fleet register Germany 1998' it is selected and the following precursor weight matrix is constructed. It contains total engine power per cell.

GRT	Length over all			
	0-10	10-20	20-50	50+
0-40	24251	54523	221	0
40-400	0	7893	39562	0
400+	0	0	0	28865

In this example data content is concentrated in part of the cells. This is generally the case in this type of integration process. This is exactly the reason why experts apply this method. The final result is in most cases rather insensitive for uncertainty in the proportionality model. The model uncertainty is usually propagated to the final result a relatively small amount of the data.

The normalized weight matrix W is obtained by dividing cell values by column totals. Multiplication WC , where C is the primary source data content, subsequently results in the goal table content:

GRT class (tons)	total catch (tons)
0-40	32970
40-400	57653
400 or more	116594

6 DISCUSSION

This paper addresses the socio-economically relevant problem of statistical information integration for political decision making. The problem occurs in countless areas of application over all levels of government and management. From a research perspective this problem is close to the areas of intelligent information integration and statistics. From another perspective the problem can be seen as introducing statistics to the intelligent information integration, which introduces a number of specific problems.

Literature study reveals that the results of intelligent information integration do not cover the specific problems of statistical information integration. An exception is [8] in which an overall model was proposed, that is dedicated to the statistical integration process used to support the European Common Fisheries Policy. That model does not use either a generic ontology of statistics, or generic models of statistical methods. Furthermore, the problem of possible classification differences was solved in ad hoc manner for specific data sources.

Statistical techniques that are an obvious source of inspiration are not generally applicable. This is caused by the inaccessibility of data and by lack of domain specific statistical models. Formalisation of human expert knowledge did not solve these problems. However, the acquired heuristic knowledge did enable the formalisation

and implementation of a model that is more generally applicable. The model uses heuristics for selecting primary sources, proportionality models, and weight matrices to overcome the classification differences of heterogeneous sources. The “weight matrix method” distilled from the expertise of humans in the field takes a central place in the model. The structure of the model is set up in such a way that it allows easy extension with other methods.

In the research reported in this paper, the focus was on heuristics and software support. From the examples that were studied, the weight matrix method appears to give reliable results. Current research is focused on extension of the system with statistical techniques to quantify the reliability of the results produced by the weight matrix method. Future research will create dedicated software for statistical techniques to further support integration of heterogeneous statistics.

REFERENCES

1. http://europa.eu.int/comm/fisheries/doc_et_publ/green1_en.htm.
2. Economic Performance of Selected European fishing Fleets, Annual report 2000, Concerted Action FAIR PL97-3541, ISBN 90-5242-624-4. LEI, Den Haag, 2000.
3. Arens Y., C.Y. Chee, C-N Hsu, C.A Knoblock: Retrieving and Integrating Data from Multiple Information Sources. *International Journal on Intelligent and Cooperative Information Systems*, 2, 1993.
4. Goh, C., S. Bressan, S. Madnick, M. Siegel: Context Interchange: New Features and Formalism for the Intelligent Integration of Information. MIT-Sloan Working Paper 3941, 1997.
5. Subrahmanian V.S., S. Adali, A. Brink, R. Emery, J.J. Lu, A. Rajput, T.J. Rogers, R. Ross, C. Ward: HERMES: Heterogeneous Reasoning and Mediator System. <http://www.cs.umd.edu/projects/hermes/publications/postscripts/tois.ps>, 1996.
6. Nodine, M., J. Fowler, T. Ksiezyk, B. Perry, M. Taylor, A. Unruh: Active Information Gathering in InfoSleuth. *International Journal on Cooperative Information Systems*, 9, 2000.
7. Knoblock, C.A., S. Minton, J.L. Ambite, N. Ashish, I. Muslea, A.G. Philpot, S. Tejada: The Ariadne Approach to Web-based Information Integration. *International Journal on Cooperative Information Systems*, 10, 2001.
8. Klinkert, M., Treur, J., Verwaart, D.: Knowledge-Intensive Gathering and Integration of Statistical Information on European Fisheries. In: R. Loganantharaj, G. Palm and M. Ali (eds.), *Proceedings IEA/AIE 2000. Lecture Notes in AI*, vol. 1821, Springer Verlag, 2000.
9. Welvaert, M.: De Belgische zeevisserij – aanvoer en besomming. Ministerie van Landbouw – Bestuur der Economische Diensten – Dienst voor de zeevisserij, Brussels, 1993.
10. Uitkomsten van de Belgische zeevisserij 1993. Ministerie van Landbouw – Bestuur der Economische Diensten – Dienst voor de zeevisserij, Brussels, 1993.
11. Wooldridge, J.M.: *Introductory Econometrics*. South-Western College Publishing, 2000.
12. Brazier, F.M.T., C.M. Jonker, J. Treur, Principals of Compositional Multi-agent Systems Development. In: J. Cuenca (ed.), *Proceedings of the 15th IFIP WCC, Conference on Information Technology and Knowledge Systems, IT&KNOWS'98*, IOS Press, 1998.