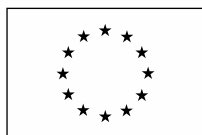


RESEARCH IN OFFICIAL STATISTICS

1 ■ 2002



An international journal for research in official statistics

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (<http://europa.eu.int>).

Luxembourg: Office for Official Publications of the European Communities, 2003

ISSN 1023-098X

© European Communities, 2003

Research in Official Statistics

ROS — An international journal for research in official statistics

ROS — Volume 5 — Number 1 — 2002

Contents

Articles

Statistical database modelling and compatibility for processing and publication in a distributed environment 5

Bryan Scotney, John Dunne, Sally McClean

Mining spatial association rules in census data 19

Donato Malerba, Floriana Esposito, Francesca A. Lisi, Annalisa Appice

Experiences in developing a spatio-temporal information system 45

Giuseppe Sindoni, Stefano De Francisci, Mario Paolucci, Leonardo Tininini

Forum

Towards standardisation of survey outcome categories and response rate calculations 61

Peter Lynn, Roeland Beerten, Johanna Laiho and Jean Martin

Effects of interviewer's workload on the Dutch Labour Force Survey 85

Jan A. Van Den Brakel

Statistical research at Statistics Norway 105

*Johan Heldal, Jan Bjørnstad, Anne Gro Hustoft, Dag Roll-Hansen,
Dinh Q. Pham and Li-Chun Zhang*

Statistical database modelling and compatibility for processing and publication in a distributed environment

Bryan Scotney (*), John Dunne (**) and Sally McClean (*)

(*) *University of Ulster, Cromore Road, Coleraine BT52 1SA, Northern Ireland*
E-mail: bw.scotney@ulst.ac.uk; si.mcclean@ulst.ac.uk

(**) *Central Statistics Office, Ireland*
Databank Administrative Unit, Skehard Road, Cork, Ireland
E-mail: dunnejo@cso.ie

Keywords: multi-data sources integration and systematisation, data models, metadata

Abstract

We consider the interoperability of information systems within a distributed environment, such as across statistical organisations of the Member States of the European Union. Within a logical layer between the physical storage of the data (server) and the presentation of the data to the user (client), we introduce a conceptual data model that can facilitate interoperability between a number of different data models and structures in a distributed environment. For interoperability, we discuss a micro–macro–metadata model that can interface readily with data warehouses and other cube object models and can exploit the accommodation of the operational metadata necessary for statistical query processing.

1. Introduction

In a ‘vision for the future’ for statistical organisations, Sundgren [9] proposes that information systems architectures of a statistical organisation will consist of survey processing systems, a corporate data warehouse and analytical processing systems. Such a data warehouse will contain compartments for raw data and metadata, final observation registers, final multidimensional tables, electronic documents and global metadata.

In this paper, we are concerned with the interoperability of such systems within a distributed environment, such as across the Member States of the European Union. Information system developers often employ a three-tier architecture in client–server applications to insert a logical layer or concept model between the physical storage of the data (server) and the presentation of the data to the user (client). Employing such a layer has many advantages. Application development is separated into components, facilitating changes to be made independently to either the presentation environment or the physical storage layer. Of principal interest in this paper is the capacity to introduce a conceptual data model into the logical layer that can facilitate interoperability between a number of different data models and structures in a distributed environment. The key is to develop a logical model that is flexible enough to accommodate data and metadata stored in a range of physical layers and can facilitate users to extract components in a variety of formats for presentation and publication. The goal is to provide a federated but unified

framework for the publication of statistics. The vision is of a system that allows suppliers of statistics to subscribe to an integrated network of data stores, whilst retaining control over access to their own data, enabling independent organisations to publish their own data within a framework that makes comparison and harmonisation possible. Since the distributed models are shown to be compatible at a high level, it would be possible to build an overarching interface for data importation, registration and processing. Such an interface could then be used to access and process in one system, data that has been imported and registered in another. However, different low-level interfaces have already been extensively developed for existing systems and have been integrated into the production processes for official statistics. Providers and producers of official statistics require the extended functionality offered by transformation between models without the burden of an additional operating environment. The key, therefore, is to facilitate interoperability of systems so that users can seamlessly work with data that originate in a range of other systems, just as if the data were registered in their own system.

2. Micro–macro–metadata models

As a mechanism for implementing interoperability we discuss a micro–macro–metadata (Mimamed) model that has been adopted and developed in the Addsia (**A**ccess to **d**istributed **d**atabases for **s**tatistical **i**nformation and **a**nalysis) and Mission (**M**ulti-agent **i**ntegration of **s**hared **s**tatistical **i**nformation **o**ver the (Inter)**n**et) projects, within the DOSIS and EPROS programmes respectively. This model supports a number of operators for analytical processing, and accommodates both raw (micro)data and aggregate (macro)data, and a variety of metadata. Much of the data constitutes final observation registers. Most of the query execution within the developed and developing systems involves macrodata, and the concept of final multidimensional tables is at the heart of the systems both for computation and publication. Global metadata is also essential to facilitate the merging and integration of data from distributed sources.

The MIMAD (micro–macro–data) model, previously developed at the University of Ulster [2], [3], [4], [5] and as discussed in [6], was extended in Addsia to a Mimamed (micro–macro–metadata) model to take account of the active metadata that is involved in statistical query execution. In Addsia, such metadata is involved in computational procedures on the object data and is stored as relational tables alongside the corresponding micro- or macro-relations in what is termed a ‘solar architecture’. In this way, the unit of summary statistical data (incorporating macro-data and metadata) for which operators (for statistical query execution) are developed is the macro–metadata object (Mameob). Prior to this extension, the data objects considered were the micro- and macro-relations, termed MIOB (microdata object) and MAOB (macrodata object) respectively.

Mameobs are defined to form the basis of a framework for the efficient implementation of aggregate statistical operators in a distributed environment. In the Addsia Mimamed model, the MAOB as originally introduced by Sadreddini et al. [2], [3], [4], [5] is partitioned into a number of summary tables, each containing only one numerical attribute. Each of these summary tables is itself a compressed form of a MAOB in the original sense. These are satellite tables of a central ‘sunkey’ table that contains a key for the categorical attribute value label combinations. An

advantage of this model is that all MAOBs can be guaranteed to have the same structure. Additionally, to avoid duplication, each summary table contains only the key attribute from the ‘sunkey’ table, rather than the categorical attributes. This is the sense in which the satellite tables are ‘compressed’ MAOBs.

A Mameob is a combination of a sunkey table and a set of compressed MAOBs. In each Mameob, therefore, the centre is a sunkey table. It contains an index for category attributes. Around it, there may be any number of MAOB summary tables, each of which contains summary data for a single numerical attribute. A dictionary table is also associated with the sunkey table. This dictionary contains information on the structure of the entire Mameob; this structure is shown in Figure 1.

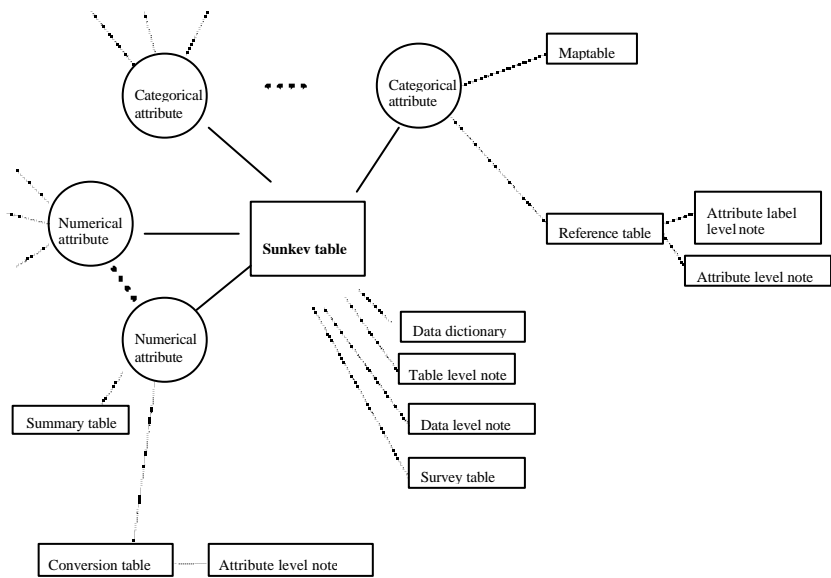


Figure 1: ‘Solar’ architecture

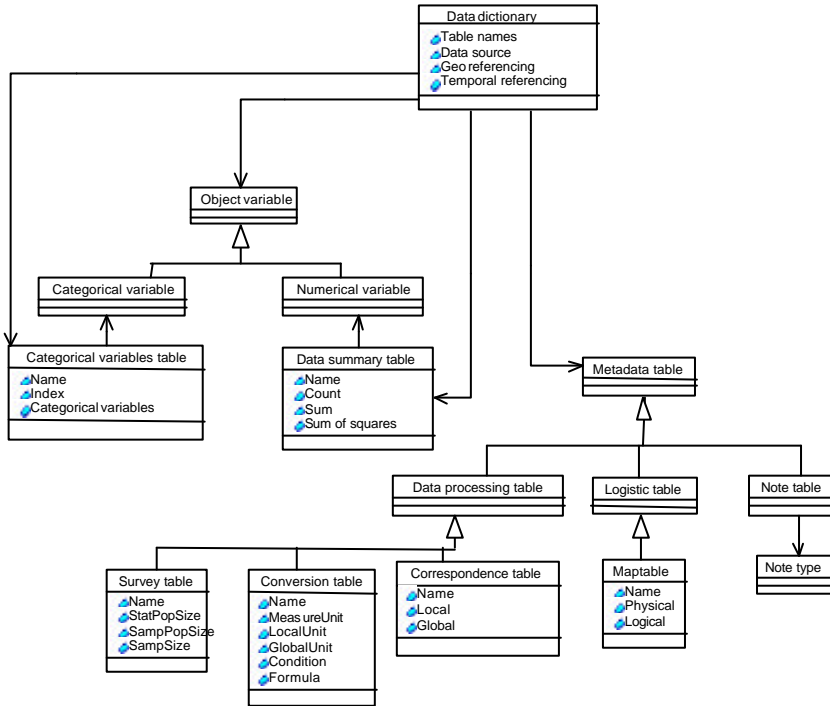


Figure 2: Elements of the Mimamed model

The Addisia Mimamed model extends the original MIMAD model through the inclusion of metadata tables associated with the main tables (i.e., the sunkey table and the summary tables). These metadata tables are of a number of different forms: attached to the sunkey table may be correspondence tables, survey tables, maptables, and note tables (containing a variety of types of note); attached to each of the summary tables may be conversion tables and note tables. These metadata tables store three types of metadata: information needed for data processing; information on logistical mappings; notes for statistical user interpretation.

Data processing information may be: correspondences between classifications for categorical variables; conversions between units for numerical variables; information about a survey. Notes may be about context, data or metadata; notes about data may be attached to individual cells, variable values, variables or tables; notes about metadata may describe correspondences between classifications or conversions between measurement units. The relationships between the elements of the Mimamed model are shown in Figure 2.

In the solar architecture, only the summary tables are macro-relations (MAOBs). A summary table, $R_V \langle \text{Sunkey}, S_1, \dots, S_m \rangle$, describes a set of m summary attributes S_1, \dots, S_m that summarise an underlying numerical attribute V in relation to a key defined by the parameter ‘sunkey’. A sunkey table, $R_keyTable \langle \text{Sunkey}, C_1, \dots, C_n \rangle$, is a central meta-relation that defines keys for a set of n categorical attributes C_1, \dots, C_n . Each of the other tables depicted is a meta-relation. Associated with every sunkey table and summary table is a set of meta-relations. These include: correspondence tables that define the relationship between local and global classification schemes ($R_C_Corr \langle \text{Local}, \text{Global} \rangle$); conversion tables that define the conversion between the

local units of V and the units used in a global ontology ($R_V_Conv \langle \text{MeasureUnit, LocalUnit, GlobalUnit, Condition, Formula} \rangle$); survey tables from which sampling and responding fractions may be derived to adjust the sample summaries to estimated population summaries (e.g. if consideration is restricted to simple unstratified random sampling, a survey table is defined by the relation $R_Survey \langle \text{StatPopSize, SampPopSize, SampSize} \rangle$); maptables that describe the mappings between logical labels and physical labels ($R_maptable \langle \text{Physical, Logical} \rangle$); note tables that contain notes relating to R as a whole (table level), to a subset of attributes in R (attribute level), to a subset of tuples in R (attribute value label level), or to a subset of attribute values in R (cell level) ($R_Note \langle \text{Name, } C_1, \dots, C_n, V_summary, \text{Text, Importance, Type} \rangle$). The elements of a *Mameob* are illustrated in Tables 1 to 7 using an example with categorical variables ‘employ’ and ‘gender’, and numerical variables ‘income’ and ‘tax’. The sunkey table is called *Employment_keyTable*:

Table 1: Employment_keyTable

Sunkey	Gender	Employ
1	M	FT
2	F	FT
3	M	PT
4	F	PT
5	M	U
6	F	U

The two summary tables are for numerical variables ‘income’ and ‘tax’ respectively:

Table 2: Employment_income

Sunkey	Income_N	Income_S	Income_SS
1	20	800	32 000
2	40	600	26 000
3	60	500	20 000
4	40	450	17 000
5	20	300	11 000
6	60	250	10 000

Table 3: Employment_tax

Sunkey	Tax_N	Tax_S	Tax_SS
1	19	200	2 000
2	40	350	3 500
3	58	300	3 200
4	40	220	1 800
5	18	50	400
6	60	110	550

We have a correspondence table between local and global classifications of employment:

Table 4: *Employment_Employ_Corr* <Local_employ, Global_employ>

Local_employ	Global_employ
*	*
FT	Temporary_FT
FT	Permanent_FT
PT	Temporary_PT
PT	Permanent_PT
U	Unemployed

and a conversion table between currencies for income:

Table 5: *Conversion table: Employment_income_conv*

MeasureUnit	LocalUnit	GlobalUnit	Condition	Formula
1 000	UK pound	euro	Year = 1988	$G \times 1.412$

A survey table provides sampling information:

Table 6: *R_Survey* <StatPopSize, SampPopSize, SampSize>

StatPopSize	SampPopSize	SampSize
1 026	530	326

and a maptable for the variable ‘employ’ relates physical and logical information:

Table 7: *Maptable R_maptable*<Physical, Logical>

Physical	Logical
P	Part time
F	Full-time
S	Student
U	Unemployed
O	Other

3. Data warehouses

Statistical analysis in NSIs falls into two broad categories: standard, regular production of pre-determined views (that will eventually form part of a regular publication) and extraction of interesting information from raw data; this may be referred to as knowledge discovery and obtained via a process of data mining. However, as the concentration of regular statistical production in terms of paper publications becomes increasingly displaced by automated production means that allow remote users to compile statistical summaries electronically, the use of OLAP (online analytical processing) in NSIs to produce regular statistical output will

increasingly replace more traditional activities [12]. Hence NSIs are likely to adopt data warehouses for use in regular statistical production as well as for knowledge discovery. OLAP databases may therefore be viewed as a third category of future statistical end-products.

In data warehouse terminology, categorical attributes are referred to as the ‘dimensions’ and numerical attributes as the ‘facts’. Storing a data cube in a relational database requires the creation of a table for all of the dimension keys of the cube and the corresponding cell values of the numerical attribute(s). This table is referred to as the ‘fact table’. For each dimension, there is a ‘dimension table’, which provides a key for each value label at the lowest ‘level of measurement’ for the dimension. In general, a dimension table is a relation of the form *Dimension_Table_Name* <key, level1,..., leveln>, where level1,..., leveln are the n levels at which the dimension may be measured. In general, a fact table is a relation of the form *Fact_Table_Name* <ID, C1_key,..., Cn_key, V_numerical>, where the keys C1_key,..., Cn_key are the key values for the n categorical variables, or dimensions that are required to describe the distribution of the numerical attribute V_numerical. If aggregation is carried out on a regular basis, it is efficient to store fact tables containing aggregates. In general, an aggregate fact table is a relation of the form *Aggregate_Fact_Table_Name* <C1_key,..., Cn_key, V_S1,..., V_Sm>, where the keys C1_key,..., Cn_key are the key values for the n categorical variables, or dimensions, that are required to describe the distribution of the m summary functions V_S1,..., V_Sm applied to the numerical attribute V.

As an example, suppose that the categorical attribute, or dimension, ‘location’ may be measured at three different levels: State, region, town. Then the dimension table ‘location’ will be of the form in Table 8. If a second dimension is ‘employment’, with dimension table as in Table 9, then a fact table for the amount of tax paid by individuals might be of the form shown in Table 10. The fact table shown in Table 10 is a micro-relation, i.e. it contains individual level raw data. If aggregation is carried out on a regular basis, it is efficient to store fact tables containing aggregates, and an example of an aggregate fact table derived by aggregation of tuples in the fact table shown in Table 10 is presented in Table 11.

Table 8: Dimension_table_location <Key, State, Region, Town>

Key	State	Region	Town
001	UK	East Anglia	Cromer
002	UK	East Anglia	Lowestoft
003	UK	Northumbria	Hexham
004	Ireland	South-East	Waterford
...

Table 9: Dimension_table_employment <Key, Employ>

Key	Employ
1	FT
2	PT
3	U

Table 10: *Fact_table_tax* <ID, Location_key, Employment_key, Tax_paid>

ID	Location_key	Employment_key	Tax_paid
0001	02	3	0
0002	02	1	10 000
0003	03	2	3 200
0004	04	2	2 800
0005	01	3	400
0006

Table 11: *Aggregate_fact_table_tax* <Location_key, Employment_key, Tax_S>

Location_key	Employment_key	Tax_S
01	1	4 456 647
01	2	876 322
01	3	20 060
02	1	3 325 600
02	2	748 285
02	3	14 387
...

Current data warehouses do not readily accommodate the structured operational metadata that is required for statistical query processing. However, it is imperative to recognise that an increasing number of NSIs envisage the adoption of data warehouses in the future, and pilot projects already exist in this area, focusing on the transition from the more traditional ‘product’ based approach to a ‘process’ approach to producing statistical outputs. For example, Statistics Netherlands is developing a new object model for statistical information in tables through two international research projects: the ‘Cube object model’ project of the Statistical Open Source (SOS) consortium, and the ‘Flexible access to statistics, tables and electronic resources’ (Faster) project. The StatLine system, a standard for electronic statistical publication based on an internal table format, is currently being revised in favour of a system based on the ‘cube object model’ [11], where the dimensions are classified into classification items, within levels, within hierarchies. Therefore, for interoperability, we propose a Mimamed model that can both interface readily with data warehouses and exploit the accommodation of the operational metadata necessary for statistical query processing accomplished in the Addsia and Mission projects.

4. A hybrid Mimamed model

In order to interface with macro-data in a data cube, a hybrid model is proposed in which a data cube may be transformed into a Mameob. The transformation takes place between the aggregate fact tables and the dimension tables in the data cube and the sunkey and summary tables in the Mameob. This requires a conversion between the relation sets {Aggregate_fact_table_name <C1_key,..., Cn_key, V_S1,..., Sm>, Dimension_table_name <key, level1,..., leveln>} and {R_keyTable <Sunkey, C1,..., Cn>, R_V <Sunkey, S1, ..., Sm>}. A data cube architecture is

shown in Figure 3, where it is assumed that each aggregate fact table contains summary functions of only one numerical attribute.

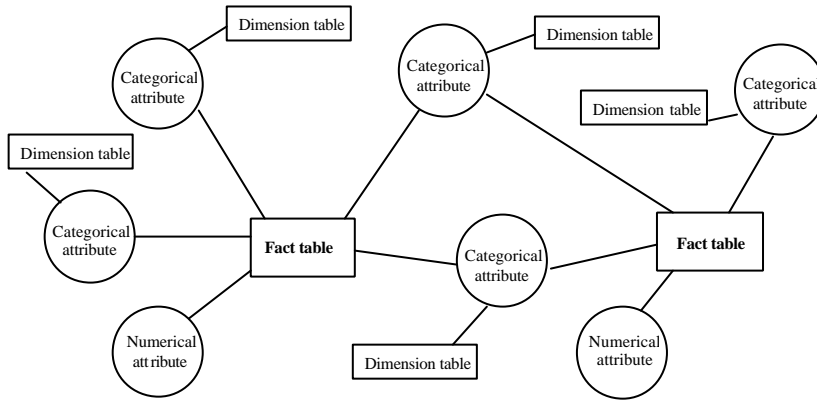


Figure 3: Data cube architecture

If there are j dimensions C_1, \dots, C_j , then the sunkey table is generated by forming a primary key (sunkey) to represent the composite key $\langle C_1_key, \dots, C_j_key \rangle$, i.e., $R_keyTable \langle Sunkey, C_1_key, \dots, C_j_key \rangle$. A particular aggregate fact table, with dimensions Ca, \dots, Cb , may then be transformed into a summary table in a Mameob by first projecting the $R_keyTable$ relation onto the dimensions Ca, \dots, Cb to produce a new keyTable ‘newkeyTable’ with a primary key ‘newsunkey’: $R_newkeyTable \langle Newsunkey, Ca_key, \dots, Cb_key \rangle$. A summary table is then produced, with the same structure as the aggregate fact table except that the composite key values $\langle Ca_key, \dots, Cb_key \rangle$ in the aggregate fact table are replaced by the corresponding key values in the newsunkey. Hence we create a summary table $R_V \langle Newsunkey, S_1, \dots, S_m \rangle$, in which $R_V.t_1.\langle S_1, \dots, S_m \rangle = aggregate_fact_table.t_2.\langle S_1, \dots, S_m \rangle$, by identifying the tuple $R_newkeyTable.t_p$ with $R_V.t_1.newsunkey = R_newkeyTable.t_p.newsunkey$ and $aggregate_fact_table.t_2.\langle Ca, \dots, Cb \rangle = \langle R_newkeyTable.t_p.\langle Ca, \dots, Cb \rangle$. Hence the creation of the summary table may be achieved by a relational table join.

As an example, consider the data cube with dimensions ‘location’, ‘gender’ and ‘employment’, and aggregate fact table `Aggregate_fact_table_tax` $\langle Location_key, Employ_key, Tax_S \rangle$. We suppose that the corresponding dimension tables and fact table are as shown in Tables 12, 13, 14 and 15 respectively.

Table 12: *Dimension_table_location* $\langle Key, Location \rangle$

Key	Location
1	UK
2	NL

Table 13: *Dimension_table_gender* <Key, Gender>

Key	Gender
1	M
2	F

Table 14: *Dimension_table_employment* <Key, Employ>

Key	Employ
1	FT
2	PT
3	U

Table 15: *Aggregate_fact_table_tax* <Location_key, Employ_key, Tax_S>

Location_key	Employ_key	Tax_S
1	1	8 856 687
2	1	5 523 600
1	2	1 462 820
2	2	897 435
1	3	462 220
2	3	188 175

First, a sunkey table is generated, as shown in Table 16. This $R_keyTable$ relation is then projected onto the dimensions ‘location’ and ‘employment’ to produce a new keyTable ‘newkeyTable’ with a primary key ‘newsunkey’, as shown in Table 17. A summary table is then produced, with the same structure as the aggregate fact table *Aggregate_fact_table_tax* <Location_key, Employ_key, Tax_S> except that the composite key values <Location_key, Employ_key> in the aggregate fact table are replaced by the corresponding key values in the newsunkey. Hence we create a summary table R_Tax <Newsunkey, S> by a relational table join in which $R_Tax.t_1.S = aggregate_fact_table.t_2.S$

by identifying the tuple $R_newkeyTable.t_p$ with

$R_Tax.t_1.newsunkey = R_newkeyTable.t_p.newsunkey$

and

$aggregate_fact_table.t_2.<Location_key, Employ_key>$

$= <R_newkeyTable.t_p.<Location_key, Employ_key>.$

This final summary table R_Tax <Newsunkey, S> is shown in Table 18.

Table 16: *R_keyTable* <Sunkey, Location_key, Gender_key, Employ_key>

Sunkey	Location	Gender	Employ
1	1	1	1
2	2	1	1
3	1	2	1
4	2	2	1
5	1	1	2
6	2	1	2
7	1	2	2
8	2	2	2
9	1	1	3
10	2	1	3
11	1	2	3
12	2	2	3

Table 17: *R_newkeyTable* <Newsunkey, Location_key, Employ_key>

Newsunkey	Location	Employ
1	1	1
2	2	1
3	1	2
4	2	2
5	1	3
2	2	3

Table 18: *R_Tax* <Newsunkey, S>

Newsunkey	Tax_S
1	8 856 687
2	5 523 600
3	1 462 820
4	897 435
5	462 220
6	188 175

5. Application

Transformation similar to that described in general for the data cube is possible to facilitate interfacing with Cristal data objects in the object model ‘Cubic, raw, or intermediate statistical data’, developed by Statistics Netherlands [10]. This form of data object may be used to access data varying from raw micro-data to multi-dimensional publication tables or cubes, all using the same structure. A Cristal data object consists of several related objects, each with a number of properties: a globally unique identifier, a name, a footnote for further explanation of the object, a description for further explanation of the object, and the time-point at which the object was last

modified. Each Cristal object contains only one statistical object (e.g. persons, companies, etc.). As in a data cube or a Mameob, a Cristal consists of a combination of classification dimensions (dimensions in a data cube; categorical attributes in a Mameob) and observation dimensions (facts in a data cube; numerical attributes in a Mameob). In a Cristal, the dimensions may be ordered into hierarchies, levels and items, just as in a data cube. The same dimension structure may be achieved in a Mameob by extension of the reference table metadata tables to include internal referencing within a classification dimension. Time and location are examples of ‘special’, or fundamental, dimensions when considering statistical analysis and, correspondingly, in all of these data structures, geographical dimensions and time dimensions are considered as special types of dimension, separate from classification dimensions.

Data in a Cristal is organised in one or more data-set objects (S_1, \dots, S_k), typically one for each survey. Each survey typically covers a number of dimensions (D_1, \dots, D_n) for a number of statistical objects (G_1, \dots, G_m) (see Figure 4). Data sets may overlap in these regards within the Cristal, providing more than one ‘datapoint’ (typically an aggregate value) for an observational dimension corresponding to a particular combination of classification dimension values. A method is therefore required to create a single data set from these different data sets for the purposes of publication (c.f. work on integration of distributed heterogeneous summary data, e.g., [1], [7], [8], and that implemented in Addsia). Furthermore, the extent of the overlap of data sets within a Cristal determines the ability to nest dimensions in a publication. For example, for a given statistical object, dimensions can be nested in a publication only if the data sets in question overlap for the statistical object and the dimensions concerned.

		D1	D2	D3	D4	D5	D6	D7	D8	D9	
Statistical objects	G1	S1					S2				
	G2										
	G3						S2				
	G4										
	G5									S4	
	G6										
	G7										

Figure 4: Data sets in a Cristal

Since only the interface to the Cristal object is defined, storage can be in any type of DBMS. It is clear that in concept the Cristal object is very similar to a MIOB or MAOB, depending on whether it contains micro-data or aggregate data. Conversion from a Cristal to a Mameob therefore requires little more than the creation of a sunkey table from the individual classification dimension keys and the creation of summary tables by the selection of individual observational dimensions.

The proposed hybrid Mimamed model should therefore readily interface with Cristal objects, whilst accommodating the operational metadata required for statistical analysis in Mission.

Within the logical layer it is therefore possible to embrace a full range of data objects, transformable between several models. Microdata Cristals (with one classification dimension and several observation dimensions) are equivalent to a simple table of raw micro-data, a fact table in a data cube having only one dimension table, or a MIOB (micro-data object); publication data Cristals (with several classification dimensions and one observation dimension) are equivalent to an aggregate fact table in a data cube, or a summary table in a Mameob. In general, Cristal objects, data cubes and MAOBs may contain several classification dimensions and several observational dimensions. In general, these data models may be made broadly equivalent by transformation, and interoperability based on the Mimamed model thus achieved. In the DCUBE project at CSO, Ireland, PC-Axis, as developed by Statistics Sweden and used by Nordic NSIs, is being investigated as the object model in the logical layer. If this proves to be successful, then transformations between the Mameob and PC-Axis formats will also be a future development.

6. Conclusions

The introduction of a conceptual data model into the logical layer of a system can facilitate interoperability between a number of different data models and structures in a distributed environment. The logical model must be flexible enough to accommodate data and metadata stored in a range of physical layers, so that users can extract components in a variety of formats for presentation and publication. The logical model is essential to providing a federated but unified framework for the publication of statistics. Providers and producers of official statistics require a system that allows suppliers of statistics to subscribe to an integrated network of data stores, whilst retaining control over access to their own data. The system must enable independent organisations to publish their own data within a framework that makes comparison and harmonisation possible. Users should not be burdened with a new operating environment that, in any case, would require external, global, maintenance: low-level interfacing between distributed data models enables autonomy of production processes within the normal operating environment of any independent organisation. Regardless of the scope of the production process, a familiar local interface is therefore enabled for any user; it is, in effect, the data sets that take on customised guises according to the particular user that is accessing them.

7. Acknowledgement

This work was partially funded by the project Mission (**M**ulti-agent **i**ntegration of **s**hared **s**tatistical **i**nformation **o**ver the (Inter)**n**et) (IST project number 1999-10655), which is part of Eurostat's EPROS initiative.

8. References

- [1] McClean, S. I., Scotney, B. W. and Greer, K. C. R., 'A scalable approach to integrating heterogeneous aggregate views of distributed databases', accepted for *IEEE trans. knowledge and data engineering*, 2002.
- [2] Sadreddini M. H., Bell, D. A. and McClean, S. I., 'Architectural considerations for providing statistical analysis of distributed data', *Information and Software Technology*, Vol. 32, 1990, pp. 459–469.
- [3] Sadreddini, M. H., Bell, D. A. and McClean, S. I., 'A model for integration of raw data and aggregate views in heterogeneous statistical databases', *Database Technology*, Vol. 4, No 2, 1991, pp. 115–127.
- [4] Sadreddini, M. H., Bell, D. A. and McClean, S. I., 'A framework for query optimisation in distributed statistical databases', *Information and Software Technology*, Vol. 34, No 6, 1992a, pp. 363–377.
- [5] Sadreddini, M. H., Bell, D. A. and McClean, S. I., 'Providing statistical functionality in a distributed environment', Westlake, A., Banks, R., Payne, C. and Orchard, T. (eds), *Survey and statistical computing*, North Holland, 1992b, pp. 467–476.
- [6] Scotney, B. W. and McClean, S. I., 'Using database technology to facilitate statistical analysis of distributed data', *New techniques and technologies for statistics II*, IOS Press, Amsterdam, 1997, pp. 203–213.
- [7] Scotney, B. W., McClean, S. I. and Rodgers, M. C., 'Optimal and efficient integration of heterogeneous summary tables in a distributed database', *The Journal of Data and Knowledge Engineering*, Vol. 29, 1999, pp. 337–350.
- [8] Scotney, B. W. and McClean, S. I., 'Efficient knowledge discovery through the integration of heterogeneous data', *Information and Software Technology*, Vol. 41, 1999, pp. 569–578.
- [9] Sundgren, B., 'An information systems architecture for national and international statistical organisations', April 1997.
- [10] Van Bracht, E., de Jonge, E. and Kaper, E., 'Cristal data objects — An object model for cubic, raw, or intermediate statistical data', Statistics Netherlands, March 2000.
- [11] Van Bracht, E. and Sluis, W., 'Towards an international standard for multi-dimensional tables', Statistics Netherlands, June 2000.
- [12] Vuscan, M., 'The application of data warehouse techniques in a statistical environment', seminar on integrated statistical information systems and related matters (ISI 2000), Riga, Latvia, May 2000.

Mining spatial association rules in census data

Donato Malerba, Floriana Esposito, Francesca A. Lisi and Annalisa Appice

*Dipartimento di Informatica, Università degli Studi di Bari,
Via Orabona, 4, I-70126 Bari*

*E-mail: malerba@di.uniba.it; esposito@di.uniba.it; lisi@di.uniba.it;
appice@di.uniba.it*

Abstract

In this paper we propose a method for the discovery of spatial association rules, that is, association rules involving spatial relations among (spatial) objects. The method is based on a multi-relational data mining approach and takes advantage of the representation and reasoning techniques developed in the field of inductive logic programming (ILP). In particular, the expressive power of predicate logic is profitably used to represent spatial relations and background knowledge (such as spatial hierarchies and rules for spatial qualitative reasoning) in a very elegant, natural way. The integration of computational logics with efficient spatial database indexing and querying procedures permits applications that cannot be tackled by traditional statistical techniques in spatial data analysis. The proposed method has been implemented in the ILP system SPADA (spatial pattern discovery algorithm). We report the preliminary results of the application of SPADA to Stockport census data.

1. Background and motivation

Censuses make a huge variety of general statistical information on society available to both researchers and the general public. Population and economic census information is of great value in planning public services (education, funds allocation, public transportation), as well as in private businesses (locating new factories, shopping malls or banks, as well as marketing particular products).

The application of data mining techniques to census data, and more generally, to official data, has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society [29]. Nevertheless, it is not straightforward and requires challenging methodological research, which is still in the initial stages.

As an illustrative example of some research issues, let us consider the census data table reported in Figure 1, where each row represents an enumeration district (ED), the smallest areal unit for which census data are published in UK ⁽¹⁾.

⁽¹⁾ National statistics institutes (NSIs) make a great effort to collect census data, but they are not the only organisations that analyse them: data analysis is often done by different institutes. By law, NSIs are prohibited from releasing individual responses to any other government agency or to any individual or business enterprise, so data are summarised for reasons of privacy before being distributed to external agencies and institutes. Therefore, data analysts are confronted with the problem of processing data which summarise characteristics of groups of individuals.

c1	c24	c25	c26	c27	c28	c30	c32	c33	c34	c35	c36
03BSFA01	44	69	23	6	5	7	0	0	7	15	109
03BSFA02	56	108	36	8	11	22	0	2	12	27	233
03BSFA03	74	98	27	5	9	18	1	0	13	33	127
...

c1: ED level code, e.g. ‘03BSFA01’, where ‘03’ denotes a country/region (Greater Manchester), ‘BS’ denotes a district (Stockport), ‘FA’ denotes a ward (Bredbury) and ‘01’ is the enumeration district.
c24: Total females of employees (full time) aged 16 and over
c25: Total males of employees (full time) aged 16 and over
c26: Total females of employees (part time) aged 16 and over
c27: Total males of employees (part time) aged 16 and over
c28: Total females of self-employed — with employees aged 16 and over
c30: Total males of self-employed — with employees aged 16 and over
c32: Total females on a government scheme aged 16 and over
c33: Total males on a government scheme aged 16 and over
c34: Total females of unemployed aged 16 and over
c35: Total males of unemployed aged 16 and over
c36: Total car availability in all households (households with three or more cars counted as having three cars)

Figure 1: An example of census data table. Data are summarised per enumeration district (ED)

The data analyst might be interested in finding some kind of dependence between the active population and the percentage of cars per household. A dependence can be expressed as an association rule, that is an implication of the form

$$P \rightarrow Q \text{ (s \% , c \%)},$$

where P and Q are a set of literals, called items, such that $P \cap Q = \emptyset$, while the percentages $s\%$ and $c\%$ are respectively called the support and the confidence of the rule, meaning that in $s\%$ of table rows both P and Q are true, and in $c\%$ of rows if P is true Q also holds. More formally, s estimates the probability $p(P \cup Q)$, while c estimates the probability $p(Q|P)$. The following is an example of an association rule establishing a dependence between the active population and the percentage of cars per household:

$$\{\text{low \%FTEM, low \%PTEM}\} \rightarrow \{\text{low \%PTEF, low \%CH}\} \quad (41\%, 62\%),$$

where low %FTEM, low %PTEM, low %PTEF, and low %CH are some items obtained by normalising and then discretising the attributes in Figure 1, namely:

low %FTEM: low (0.. 34 %) percentage of full-time employed males

low %PTEM: low (0.. 20 %) percentage of part-time employed males

low %PTEF: low (0.. 16 %) percentage of part-time employed females

low %CH: low (0.. 0.8 %) percentage of cars per household

For the sake of completeness, we report an alternative logical notation for the above association rule:

$\text{low \%FTEM} \wedge \text{low \%PTM} \rightarrow \text{low \%PTEF} \wedge \text{low \%CH} (41 \%, 62 \%),$

where the conjunction $\text{low \%FTEM} \wedge \text{low \%PTM} \wedge \text{low \%PTEF} \wedge \text{low \%CH}$ is called pattern. This association rule states that in 62 % of EDs where there is both a low percentage of full-time employed males and a low percentage of part-time employed males, the percentage of part-time employed females is low and the percentage of cars per household is also low. The support is 41 %, meaning that in 41 % of analysed EDs all conditions expressed by the pattern $\text{low \%FTEM} \wedge \text{low \%PTM} \wedge \text{low \%PTEF} \wedge \text{low \%CH}$ holds. By interpreting this association rule we can say that 41 % of EDs seem to be deprived areas.

1.1. The single table assumption

The discovery of association rules has attracted a great deal of attention in data mining research [11]. The blueprint for all the algorithms proposed in the literature is the levelwise method by Mannila and Toivonen [24], which is based on a breadth-first search in the lattice spanned by a generality order between patterns. Despite some interesting extensions, almost all algorithms reported in the literature share a restrictive data representation formalism, known as single-table assumption. More specifically, it is assumed that the data to be mined are represented in a single table (or relation) of a relational database, such that each row (or tuple) represents an independent unit of the sample population and the columns correspond to properties of units.

In some applications this assumption turns out to be a great limitation. For instance, in the above example, units correspond to EDs, which are spatial objects, since they have a geographical location. Having recognised this peculiarity, the data analyst may be interested in investigating the socioeconomic phenomenon of deprivation in association with the geographical distribution of EDs. To achieve this goal, the analyst may decide to augment the data table in Figure 1 with information on neighbouring units. In particular, for each ED in Figure 1, the analyst proposes the following data specifications:

- the number of schools in the neighbouring EDs,
- the number of banks in the neighbouring EDs, and
- the number of commercial activities in the neighbouring EDs,

since he/she suspects that the low percentage of cars can also be related to the number of services available in the neighbourhood.

If the analyst decides to represent the above data only for one neighbouring ED, the data table in Figure 1 can be extended by simply adding three attributes (see Table 1). What if he/she wants to represent the three attributes for all spatially adjacent EDs, which are variable in number? Under the single-table assumption he/she can create one entry for each adjacent ED in the original data table (see Table 2). However, this solution presents two main disadvantages:

- (i) we have the usual problems connected with non-normalised tables, such as redundancy and anomalies in the insertion and removal of data.
- (ii) we have one line per neighbouring ED, which means that the analysis results will really concern neighbouring EDs. In other words, the observation unit has deceptively changed.

Table 1: Three additional attributes of the nearest neighbour added to the single table

c1	c24	c25	...	c36	Number of schools	Number of banks	Number of commercial activities
03BSFA01	44	69	...	109	1	1	13
03BSFA02	56	108	...	233	0	0	23
03BSFA03	74	98	...	127	0	1	6
...

Table 2: Each row in the original table is duplicated to add information on a neighbouring ED

c1	c24	c25	...	c36	Number of schools	Number of banks	Number of commercial activities
03BSFA01	44	69	...	109	0	1	2
03BSFA01	44	69	...	109	1	0	3
03BSFA03	74	98	...	127	0	1	1
...

The former is a typical database issue, while the latter is more related to the data analysis procedure.

To solve these problems and keep the single-table assumption, the data analyst may try to summarise the information on the neighbouring EDs, say, by averaging the number of schools, banks and commercial activities (see Table 3). It is noteworthy that in this case there is no redundancy and standard data mining methods work well. However, there is an information loss that might lead to the misunderstanding of the underlying phenomenon. For instance, an ED can be adjacent to another ED with many services, as well as to other EDs with no services at all, since they fall into the green belt of the city. By averaging the number of services per neighbouring ED, the analyst may give a totally wrong indication on the accessibility of services.

Table 3: Three additional attributes of the nearest neighbour added to the single table

c1	c24	c25	...	c36	Average number of schools	Average number of banks	Average number of commercial activities
03BSFA01	44	69	...	109	0.25	0.25	3.3
03BSFA02	56	108	...	233	0.33	0	0.36
03BSFA03	74	98	...	127	0	0.2	0.12
...

From a database perspective, the best representation of data would be that in Figure 2.

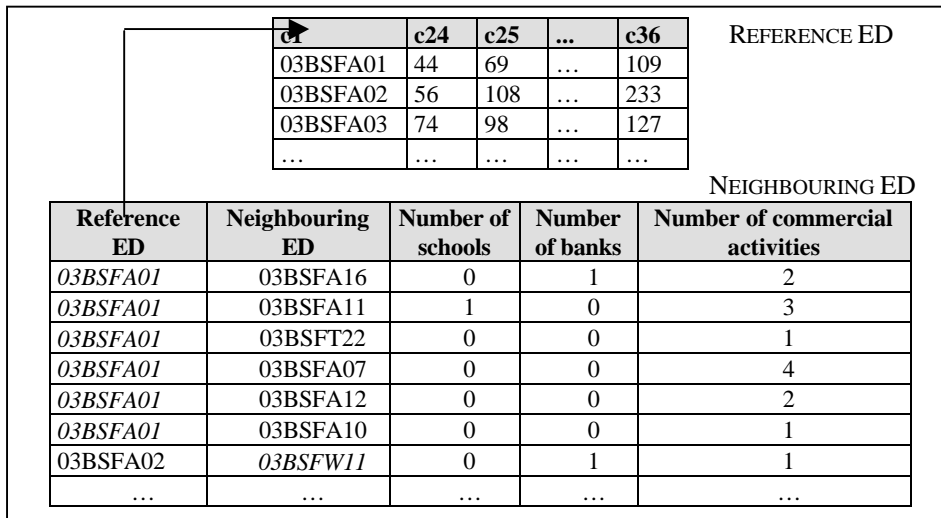


Figure 2: A multi-relation representation of socioeconomic attributes of some reference EDs and of their neighbouring ED. The attribute 'Reference ED' in the lower table is a foreign key of the upper table.

In this database two relations are defined, one for the reference EDs, that is, the EDs whose socioeconomic factors are the subject of investigation, and one for the neighbouring EDs, which are considered task relevant, because they are spatially adjacent to some reference EDs. Obviously, mining this simple database requires far more powerful methods which go beyond the single-table assumption.

1.2. A multi-relational data mining approach

The recently promoted **(multi-)relational** ⁽²⁾ approach to data mining [9] looks for patterns that involve multiple relations of a relational database. Thus the data taken as input by these approaches typically consists of several tables and not just a single one, as is the case in

⁽²⁾ The term multi-relational data mining is sometimes preferred to relational data mining and has also been used to denote data mining applied to relational databases [15].

most existing data mining approaches. Patterns found by these approaches are called **relational** and are typically stated in a more expressive language than patterns defined in a single data table.

The following is an example of a **relational association rule**:

$$\begin{aligned} &\text{male-full-time-employee\%}(X, \text{low}) \wedge \text{male-part-time-employee\%}(X, \text{low}) \wedge \\ &\text{neighbour}(X, Y) \wedge \text{comm-activities}(Y, \text{high}) \rightarrow \text{male-self-employed\%}(X, \text{high}) \\ &\hspace{15em} (32\%, 70\%), \end{aligned}$$

which states that in 70 % of the cases the low percentage of full-time and part-time male employees in some reference ED X , adjacent to another task relevant ED Y , with many commercial activities, implies a high percentage of self-employed males in X . The relational pattern

$$\begin{aligned} &\text{male-full-time-employee\%}(X, \text{low}) \wedge \text{male-part-time-employee\%}(X, \text{low}) \wedge \\ &\text{neighbour}(X, Y) \wedge \text{comm-activities}(Y, \text{high}) \wedge \text{male-self-employed\%}(X, \text{high}) \end{aligned}$$

occurs in 32 % of reference EDs.

It is noteworthy that in this example, and more generally in relational association rules, the items are first-order logic **atoms**, that is, **n -ary predicates** applied to **n terms**. In this example terms can be either **variables**, such as X and Y , or **constants**, such as **low** or **high**. In other words, subsets of **first-order logic**, which is also called predicate calculus or relational logic, are used to express relational patterns and relational association rules.

This strong link with logics is not surprising, since any relational database can be easily modelled by a **deductive relational database** (DDB), by simply transforming all tuples in materialised tables into ground facts (extensional part of a DDB) and all views into rules (intensional part of a DDB) (see Figure 3).

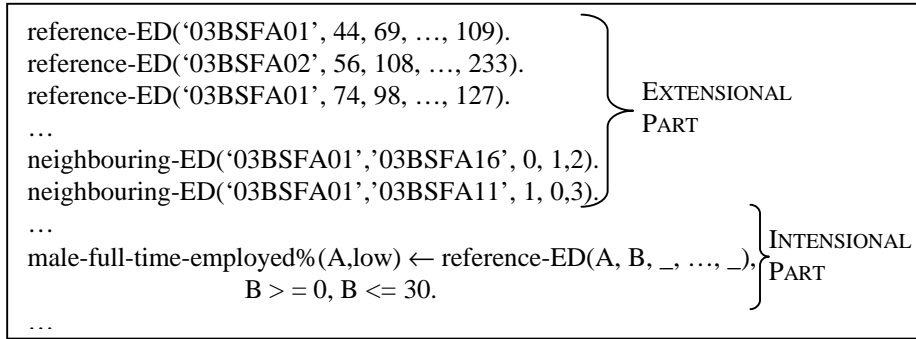


Figure 3: A deductive database view of the relational database in Figure 2. The extensional part is a set of ground facts corresponding to the tuples of the materialised tables, while the intensional part is a set of logical rules defining the views created in the relational database.

In this example, we assume that the attribute *male-full-time-employed%* has been defined by creating a view in the relational database.

Therefore, a relational pattern is simply a DDB query, whose result set cardinality corresponds to the support.

Considering this strong link with logics, it is not surprising that many algorithms for multi-relational data mining originate from the field of **inductive logic programming** (ILP) [25], [8], [19], [26]. ILP has always been concerned with finding patterns expressed as logic programs. Initially, its main focus was on automated program synthesis from examples [2], but, in recent years, the scope of ILP has broadened to cover the whole spectrum of data mining tasks (association rules, regression, clustering and so on).

Extending a single-table data mining algorithm to a relational one is not trivial. Considerable insight and creativity is required to extend some key notions, such as distance measure and probabilistic dependence, to multi-relational data. Efficiency is also very important, as even testing a given relational pattern for validity is often computationally expensive. Moreover, for relational pattern languages, the number of possible patterns can be very large and it becomes necessary to limit their space of possible patterns by providing explicit constraints (**declarative bias**). These normally specify what relations should be involved in the patterns, how the relations may be interconnected and what other syntactic constraints the patterns have to obey.

1.3. Additional issues in spatial data mining

As explained above, there are two reasons for approaching the problem of mining spatial association rules as a multi-relational data mining problem. First, attributes of the neighbours of some spatial object of interest may influence the object itself, hence the need for representing object interactions. Second, different geographical objects may have

different properties, which can be properly modelled by as many data tables as the number of object types, hence the inadequacy of the single-table representation.

Some proposals for mining **relational** association rules have already been reported in literature [6]. However, mining **spatial** association rules is a more complex task. Two further degrees of complexity are:

- (i) the implicit definition of spatial relations and
- (ii) the granularity of the spatial objects.

The former is due to the fact that the location and the extension of spatial objects **implicitly** defines spatial relations such as topological, distance and direction relations. Therefore, complex data transformation processes are required to make spatial relations explicit (see the application of machine learning techniques to topographic map interpretation [22]).

The latter refers to the fact that spatial objects can be described at multiple levels of granularity. For instance, UK census data can be geo-referenced with respect to the following hierarchy:

ED → Ward → District → County,

based on the **inside** relationship between locations ⁽³⁾. Interesting rules are more likely to be discovered at low granularity levels (ED and ward) than at county level. On the other hand, large support is more likely to exist at higher granularity levels (district and county) rather than at low levels.

In the next section, a new algorithm for mining spatial association rules is reported. The algorithm, named SPADA (**s**patial **p**attern **d**iscovery **a**lgorithm), is based on an ILP approach to relational data mining and permits the extraction of multi-level association rules, that is, association rules involving spatial objects at different granularity levels. SPADA has been implemented in Sictus Prolog and is interfaced to an Oracle8i® database, empowered by an Oracle Spatial cartridge, which enables spatial data to be stored, accessed and analysed quickly and efficiently. The system also performs the appropriate data transformation by extracting spatial features (Featex module) and by discretising numerical attributes (RUDE module). The application of SPADA to two data mining tasks involving UK census data is reported in Section 3.

⁽³⁾ In particular, the Stockport district of Greater Manchester is divided into 22 wards (Bredbury, Brinnington, Cale Green, Cheadle, Cheadle Hulme North, Cheadle Hulme South, Davenport, East Bramhall, Edgeley, Great Moor, Hazel Grove, Heald Green, Heaton Mersey, Heaton Moor, Manor, North Marple, North Reddish, Romiley, Shipping, South Marple, South Reddish, West Bramhall), each of which consists of 30 EDs on average.

2. Mining spatial association rules with SPADA

The discovery of spatial association rules is a descriptive mining task aiming to detect associations between **reference objects** and some **task-relevant objects**. The former are the main subject of the description, while the latter are spatial objects that are relevant for the task in hand and are spatially related to the former. For instance, we may be interested in describing a given area by finding associations between large towns (reference objects) and spatial objects belonging to the map layers of road network, hydrography and administrative boundaries (task-relevant objects). In particular, we look for **spatial** patterns, namely patterns that contain at least one spatial relationship. We call

$$P \rightarrow Q (s \%, c \%)$$

a **spatial association rule**, if $P \cup Q$ is a spatial pattern.

As usual in the problem setting of association rule mining, we search for spatial associations with large support and high confidence (*strong rules*), such as

$$\begin{aligned} &\text{is_a}(X, \text{large_town}) \wedge \text{intersects}(X, Y) \wedge \text{is_a}(Y, \text{road}) \rightarrow \\ &\text{intersects}(X, Z) \wedge \text{is_a}(Z, \text{road}) \wedge Z \neq Y \end{aligned} \quad (91 \%, 85 \%),$$

which states that ‘**If** a large town X intersects a road Y , **then** X intersects a road Z distinct from Y **with** 91 % **support** and 85 % **confidence**’.

Since some kind of taxonomic knowledge of task-relevant geographic layers may also be taken into account to obtain descriptions at different granularity levels (**multiple-level association rules**), finer-grained answers to the above query are also expected, such as:

$$\begin{aligned} &\text{is_a}(X, \text{large_town}) \wedge \text{intersects}(X, Y) \wedge \text{is_a}(Y, \text{regional_road}) \rightarrow \\ &\text{intersects}(X, Z) \wedge \text{is_a}(Z, \text{main_trunk_road}) \wedge Z \neq Y \end{aligned} \quad (45 \%, 90 \%),$$

which provides more insight into the nature of the task relevant objects Y and Z , according to the spatial hierarchy reported in Figure 4.

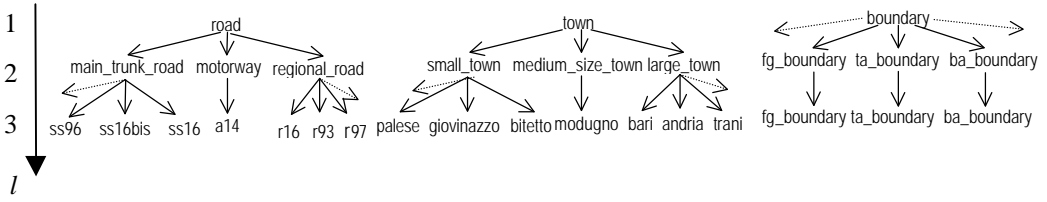


Figure 4: Three spatial hierarchies and their association to three granularity levels

It is noteworthy that the support and the confidence of the last rule changed. Generally, the lower the granularity level, the lower the support of association rules. Therefore, we follow Han and Fu's [14] proposal to use different thresholds of support and confidence for different granularity levels.

The problem of mining spatial association rules can be formally stated as follows:

Given:

- a spatial database (SDB),
- a set of reference objects S ,
- some sets R_k , $1 \leq k \leq m$, of task-relevant objects
- some spatial hierarchies H_k involving objects in R_k
- M granularity levels in the descriptions (1 is the highest while M is the lowest)
- a set of granularity assignments ψ_k which associate each object in H_k with a granularity level
- a couple of thresholds $minsup[l]$ and $minconf[l]$ for each granularity level

Find: strong multi-level spatial association rules.

The problem has been already tackled by Koperski et al. [18]. They propose a top-down, progressive refinement method which exploits taxonomies both on spatial predicates (two-step spatial computation) and spatial objects (pattern discovery). The method has been implemented in the module Geo-associator of the spatial data mining system GeoMiner [16]. This method, however, suffers from severe limitations due to the single-table assumption. Our aim is to show the usefulness of an ILP approach to mining spatial association rules and, more generally, to spatial data mining. Representation problems and algorithmic issues related to the application of our logic-based computational method are discussed in the next two subsections.

2.1. The representation

The basic idea in our proposal is that a spatial database boils down to a deductive relational database (DDB) once the spatial relationships between reference objects and task-relevant objects have been extracted. The expressive power of first-order logic in databases also

allows us to specify background knowledge (BK), such as spatial hierarchies, **constraints** on spatial patterns and association rules (**declarative bias**), as well as **domain specific knowledge** expressed as sets of **rules**. In particular, the declarative bias helps to constrain the search in the exponentially large space of patterns, so that only interesting patterns are actually generated and evaluated. On the contrary, the specification of a domain specific knowledge allows SPADA to search for patterns which could not be otherwise found in the spatial database. The rules defining the domain specific knowledge are stored in the intensional part of the DDB and can support, amongst other things, spatial qualitative reasoning. The current version of SPADA supports the specification of both the declarative bias and the domain specific knowledge, which should be considered additional input to the system.

Henceforth, we denote the DDB in hand $D(S)$ to mean that it is obtained by adding the data extracted from SDB regarding the set of reference objects S to the previously supplied BK. The ground facts in $D(S)$ can be grouped into distinct subsets: each group, uniquely identified by the corresponding reference object $s \in S$, is called **spatial observation** and denoted $O[s]$. We define the set:

$$R[s] = \{r_i \mid \exists k: r_i \in R_k \text{ and a ground fact } \theta(s, r_i) \text{ exists in } D(S)\}$$

as the set of task-relevant objects related to s . The set $O[s]$ is given by

$$O[s] = O[s|s] \cup \bigcup_{r_i \in R[s]} O[r_i | s]$$

where:

§ $O[s|s]$ contains properties of s and spatial relations between s and r_i

§ $O[r_i|s]$ contains properties of r_i and spatial relations between r_i and some $s' \in S$.

In an extreme case, $O[s]$ can coincide with $D(S)$. This is the case in which s is spatially related to all task-relevant objects. The unique reference object associated to a spatial observation allows us to define the support and the confidence of a spatial association rule. More precisely, the spatial association rule $P \rightarrow Q (s \%, c \%)$ means that in $s \%$ of spatial observations both conjunctions P and Q hold and in $c \%$ of spatial observations where P is true Q holds too. Note that the notion of spatial observation in SPADA adapts the notion of **interpretation**, which is common to many relational data mining systems [9], to the case of spatial databases.

Example 1: Suppose the mining task is to discover the associations relating large towns (S) with waterways (R_1), roads (R_2) and province boundaries (R_3) in the Province of Bari, Italy. We are also given a BK including the spatial hierarchies of interest and three levels of granularity (see Figure 4).

hierarchy(town, 1, null, [town]).

hierarchy(town, 2, town, [large_town, medium_size_town, small_town]).

```

hierarchy(town, 3, large_town, [bari, altamura, andria, barletta, trani, bitonto, molfetta,
    gravina, monopoli, corato, gioia_del_colle]).
hierarchy(town, 3, medium_size_town, [modugno, palo_del_colle, terlizzi, ruvo, noicattaro,
    adelfia, grumo, giovinazzo, mola_di_bari]).
hierarchy(town, 3, small_town, [palese, bitetto, binetto, toritto, valenzano, cassano,
    marioito, palombaio]).
hierarchy(road, 1, null, [road]).
hierarchy(road, 2, road, [motorway, main_trunk_road, regional_road]).
hierarchy(road, 3, motorway, [a14]).
hierarchy(road, 3, main_trunk_road, [ss16, ss16bis, ss96, ss98, ss99, ss100]).
hierarchy(road, 3, regional_road, [r16, r93, r97, r170, r171, r172, r271, r378]).
hierarchy(water, 1, null, [water]).
hierarchy(water, 2, water, [sea, river]).
hierarchy(water, 3, sea, [adriatico]).
hierarchy(water, 3, river, [ofanto, lacone]).
hierarchy(boundary, 1, null, [boundary]).
hierarchy(boundary, 2, boundary, [fg_boundary, ta_boundary, br_boundary, mt_boundary,
    pz_boundary]).
is_a(X, Y):- hierarchy(_, _, Y, Nodes), member(X, Nodes).
is_a(X, Y):- hierarchy(Root, _, Father, Nodes), member(X, Nodes), is_a(Father, Y).

```

Here, the **is_a** stands for an **instance_of** relation between spatial objects and their geographical layers. Spatial relations between objects in S and objects in any of R_1 , R_2 and R_3 , are extracted by means of spatial computation and transformed into facts of the kind $\langle \text{spatial relation} \rangle(\text{RefObj}, \text{TaskRelevantObj})$ to be added to $D(S)$.

Spatial observations are portions of $D(S)$, each concerning a reference object. In our case, there are 11 distinct spatial observations, one for each large town. For instance, $O[\text{barletta}]$ is given by the union of the sets of ground facts listed in Table 4. By definition, the observation encompasses not only spatial relationships between the reference object $\text{barletta} \in S$ and task-relevant objects in R_1 (adriatico etc.), R_2 (a14 etc.), R_3 (fg_boundary etc.), but also spatial relationships between each of these task-relevant objects and some other $s' \in S$ (e.g. giovinazzo) like in $\text{adjacent_to}(\text{giovinazzo}, \text{adriatico})$.

Let $A = \{a_1, a_2, \dots, a_i\}$ be a set of atoms whose terms are either variables or constants (Datalog atoms [4]). Predicate symbols used for A are all those permitted by the user-specified declarative bias, while the constants are only those defined in $D(S)$. The atom denoting the reference objects is called **key atom**. For instance, with reference to the above example of the Province of Bari, A contains the key atom $\text{is_a}(X, \text{large_town})$, ‘spatial’ atoms such as $\text{close_to}(X, Y)$, $\text{intersects}(X, Y)$, and $\text{adjacent_to}(X, Y)$, and ‘taxonomic’ atoms such as $\text{is_a}(X, \text{road})$, $\text{is_a}(X, \text{main_trunk_road})$, ..., $\text{is_a}(X, \text{water})$, $\text{is_a}(X, \text{sea})$.

Conjunctions of atoms on A are called **atomsets** [5] like the item sets in classical association rules. In our framework, a language of patterns $L[l]$ at the granularity level l is a set of well-

formed atomsets generated on A . Necessary conditions for an atom set P to be in $L[l]$ are the presence of the key atom, the presence of ‘taxonomic’ atoms exclusively at the granularity level l , the linkedness [17] and the safety. In particular, the last property guarantees the correct evaluation of patterns when the handling of negation is required (see Example 2). To a pattern P we assign an existentially quantified conjunctive formula $eqc(P)$ obtained by turning P into a Datalog query.

Table 4: The spatial observation $O[\text{barletta}]$

$O[\text{barletta} \mid \text{barletta}]$ $\text{is_a}(\text{barletta}, \text{large_town}).$ $\text{adjacent_to}(\text{barletta}, \text{adriatico}).$ $\text{Intersects}(\text{barletta}, \text{a14}).$ $\text{Intersects}(\text{barletta}, \text{ss16}).$ $\text{Intersects}(\text{barletta}, \text{ss16bis}).$ $\text{Intersects}(\text{barletta}, \text{r170}).$ $\text{Intersects}(\text{barletta}, \text{r193}).$ $\text{close_to}(\text{barletta}, \text{fg_boundary}).$...	$O[\text{a14} \mid \text{barletta}]$ $\text{is_a}(\text{a14}, \text{road}).$ $\text{intersects}(\text{bari}, \text{a14}).$ $\text{intersects}(\text{trani}, \text{a14}).$ $\text{intersects}(\text{bitonto}, \text{a14}).$ $\text{intersects}(\text{gioia_del_colle}, \text{a14}).$ $\text{intersects}(\text{molfetta}, \text{a14}).$... $O[\text{ss16} \mid \text{barletta}]$ $\text{is_a}(\text{ss16}, \text{road}).$ $\text{intersects}(\text{bari}, \text{ss16}).$ $\text{intersects}(\text{trani}, \text{ss16}).$ $\text{intersects}(\text{monopoli}, \text{ss16}).$ $\text{intersects}(\text{molfetta}, \text{ss16}).$...	$O[\text{r170} \mid \text{barletta}]$ $\text{is_a}(\text{r170}, \text{road}).$ $\text{intersects}(\text{andria}, \text{r170}).$... $O[\text{r193} \mid \text{barletta}]$ $\text{is_a}(\text{r193}, \text{road}).$... $O[\text{fg_boundary} \mid \text{barletta}]$ $\text{is_a}(\text{fg_boundary}, \text{boundary}).$ $\text{adjacent_to}(\text{trani}, \text{fg_boundary}).$... $O[\text{ss16bis} \mid \text{barletta}]$ $\text{is_a}(\text{ss16bis}, \text{road}).$ $\text{intersects}(\text{bari}, \text{ss16bis}).$ $\text{intersects}(\text{trani}, \text{ss16bis}).$ $\text{intersects}(\text{molfetta}, \text{ss16bis}).$...
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Definition: A pattern P covers an observation $O[s]$ if $eqc(P)$ is true in $O[s] \cup BK$.

Example 2: The pattern

$$P \equiv \text{is_a}(X, \text{large_town}), \text{intersects}(X, R), \text{intersects}(Y, R), Y \setminus = X, \text{is_a}(R, \text{road})$$

covers the spatial observation $O[\text{barletta}]$ shown in Table 1 because the corresponding

$$\begin{aligned} eqc(P) \equiv & \exists \text{is_a}(X, \text{large_town}) \wedge \text{intersects}(X, R) \wedge \text{intersects}(Y, R) \\ & \wedge Y \setminus = X \wedge \text{is_a}(R, \text{road}) \end{aligned}$$

is satisfied by $O[\text{barletta}] \cup BK$. Here the predicate $\setminus =$ is the ISO prolog standard built-in predicate for the non-unifiability of two variables. Note that it hides a negation.

Definition: Let O be the set of spatial observations in $D(S)$ and O_P denote the subset of O containing the spatial observations covered by the pattern P . The support of P is defined as $\sigma(P) = |O_P| / |O|$.

Definition: A spatial association rule in $D(S)$ at the granularity level l is an implication of the form

$$P \rightarrow Q \ (s\%, c\%),$$

where $P \cup Q \in L[l]$, $P \cap Q = \emptyset$, P includes the key atom and at least one spatial relationship is in $P \cup Q$. The percentages s and c are respectively called the support and the confidence of the rule, meaning that $s\%$ of spatial observations in $D(S)$ is covered by $P \cup Q$ and $c\%$ of spatial observations in $D(S)$ that is covered by P is also covered by $P \cup Q$.

Definition: The support and the confidence of a spatial association rule $P \rightarrow Q$ are given by

$$s = \sigma(P \cup Q) \text{ and } c = \phi(Q|P) = \sigma(P \cup R) / \sigma(P).$$

In multi-level association rule mining, an *ancestor* relation between two patterns at different granularity levels $P \in L[l]$ and $P' \in L[l']$, $l < l'$ exists if and only if P' can be obtained from P by replacing each spatial object $h \in H_k$ at granularity level $l = \psi_k(h)$ with a spatial object $h' < h$ in H_k , which is associated with the granularity level $l' = \psi_k(h')$.

The frequency of a pattern depends on the granularity level of task-relevant spatial objects.

Definition: Let $\text{minsup}[l]$ and $\text{minconf}[l]$ be two thresholds setting the minimum support and the minimum confidence respectively at granularity level l . A pattern P is **large** (or frequent) at level l if $\sigma(P) \geq \text{minsup}[l]$ and all ancestors of P with respect to the hierarchies H_k are large at their corresponding levels. The confidence of a spatial association rule $P \rightarrow Q$ is high at level l if $\phi(Q|P) \geq \text{minconf}[l]$. A spatial association rule $P \rightarrow Q$ is **strong** at level l if $P \cup Q$ is large and the confidence is high at level l .

2.2 Method

The task of mining spatial association rules itself can be split into two sub-subtasks:

- (i) find large (or frequent) spatial patterns;
- (ii) generate highly-confident spatial association rules.

The reason for such a division is that frequent patterns are not commonly considered useful for presentation to the user as such. They can be efficiently post-processed into rules that exceed given threshold values. In the case of association rules the threshold values of support and confidence offer a natural way of pruning weak, rare rules.

Algorithm design for frequent pattern discovery has turned out to be a popular topic in data mining. Most algorithms proposed in the literature are based on a breadth-first search in the lattice spanned by a generality order \geq between patterns. Given two patterns P_1 and P_2 , we write $P_1 \geq P_2$ to denote that P_1 is more general than P_2 or equivalently that P_2 is more

specific than P_1 . The space is searched one level at a time, starting from the most general patterns and iterating between the candidate generation and candidate evaluation phases. The high-level algorithm of SPADA implements the aforementioned levelwise method (see Figure 5).

Cycle on the level ($l \geq 1$) of the spatial hierarchies
 Find large 1-atomsets at level l
Cycle on the depth ($k > 1$) of search in the pattern space
 Generate candidate k -atomsets at level l from large $(k-1)$ -atomsets
 Generate large k -atomsets at level l from candidate k -atomsets
Until the user-defined maximum depth
Until the user-defined maximum granularity level M

Figure 5: A high-level view of the levelwise mining algorithm SPADA.

The pattern space is structured according to the θ -subsumption [28]. Many ILP systems adopt θ -subsumption as the generality order for clause spaces. In this context we need to adapt the framework to the case of atomsets. More precisely, the restriction of θ -subsumption to **Datalog queries** (i.e. existentially quantified conjunctions of Datalog atoms) is of particular interest.

Definition: Let Q_1 and Q_2 be two queries. Then Q_1 **q-subsumes** Q_2 if and only if there exists a substitution θ such that $Q_1 \supseteq Q_2\theta$.

Example 3: Let us consider the queries

$$\begin{aligned} Q_1 &\equiv \exists \text{ is_a}(X, \text{large_town}) \wedge \text{intersects}(X, R) \wedge \text{is_a}(R, \text{road}) \\ Q_2 &\equiv \exists \text{ is_a}(X, \text{large_town}) \wedge \text{intersects}(X, Y) \\ Q_3 &\equiv \exists \text{ is_a}(X, \text{large_town}) \end{aligned}$$

We say that Q_1 θ -subsumes Q_2 and Q_2 θ -subsumes Q_3 with substitutions $\theta_1 = \{Y \setminus R\}$ and $\theta_2 = \emptyset$ respectively.

We can now introduce the generality order adopted in SPADA.

Definition: Let P_1 and P_2 be two patterns. Then P_1 is more general than P_2 under θ -subsumption, denoted as $P_1 \geq_\theta P_2$, if and only if P_2 θ -subsumes P_1 .

A graphical representation of the lattice spanned by \geq_θ including the queries reported in example 3 is shown in Figure 6.

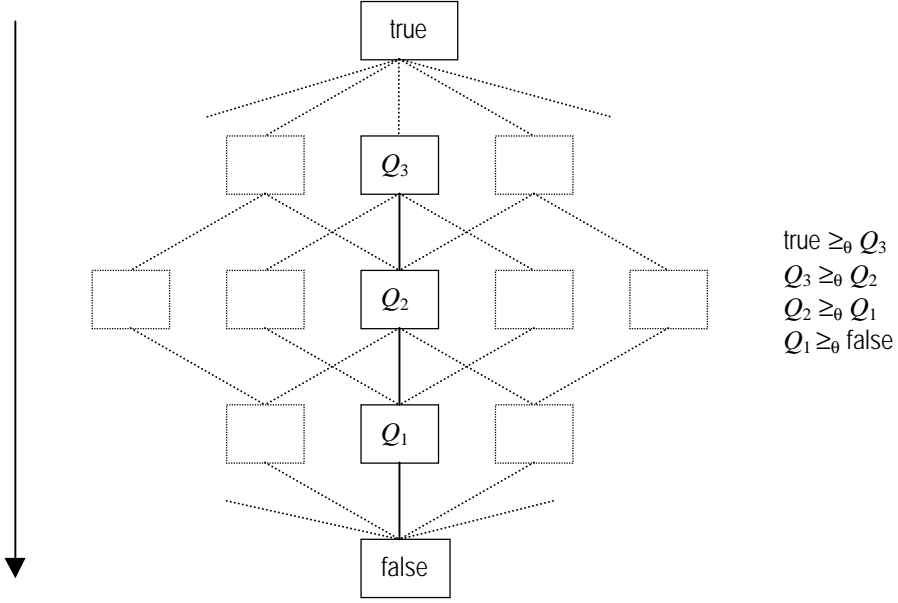


Figure 6: Example of pattern space structured according to 3_q

For θ -subsumption the following properties hold:

- reflexivity: $P \geq_\theta P$;
- transitivity: $P_1 \geq_\theta P_2$ and $P_2 \geq_\theta P_3$, then $P_1 \geq_\theta P_3$;
- decidability: a procedure exists to decide if $P_1 \geq_\theta P_2$.

The anti-symmetric property does not hold for θ -subsumption, therefore θ -subsumption is a **quasi-ordering**. It follows that, given two queries such that $P_1 \geq_\theta P_2$ and $P_2 \geq_\theta P_1$, we cannot conclude that P_1 and P_2 are equal up to renaming, i.e. P_1 and P_2 are not alphabetic variants ⁽⁴⁾. As shown below, this feature has to be taken into account during the search.

A quasi-ordered set of patterns can be searched by a **refinement operator**, namely a function which computes a set of refinements of a pattern. In particular, we need a refinement operator under θ -subsumption that enables the bottom-up search of the pattern space from the most specific to the most general patterns.

Definition: Let $\langle G, \geq_\theta \rangle$ be a pattern space ordered according to \geq_θ . A **downward refinement operator under q -subsumption** is a function ρ such that $\rho(P) \subseteq \{Q \mid P \geq_\theta Q\}$.

⁽⁴⁾ Let E and F be two expressions. Then E and F are **variants**, denoted $E \approx F$, if and only if substitutions θ and σ exist such that $E = F\theta$ and $F = E\sigma$. We also say that E is an **alphabetic variant** of F . For instance, $f(X)$ and $f(Y)$ are alphabetic variants.

It is noteworthy that \geq_θ on patterns represented as Datalog queries is monotone with respect to support, which is the criterion for candidate evaluation in SPADA. Therefore, the refinement operator drives the search towards patterns with decreasing support. Moreover, all refinements $\rho(P)$ of an infrequent pattern P are infrequent. This is the first-order counterpart of one of the properties holding in the family of the a priori-like algorithms [1], on which the pruning criterion is based.

For each granularity level (l), SPADA generates and evaluates candidates by searching the pattern space. The **candidate generation** phase consists of a refinement step followed by a pruning step. The former applies the refinement operator under θ -subsumption to patterns previously found to be frequent by preserving the property of linkedness [17]. The latter mainly involves verifying that candidate patterns do not θ -subsume any infrequent pattern. Further pruning criteria have been implemented in SPADA. In particular, the system checks that candidates are not alphabetic variants of previously discovered patterns. The complexity of this test is $O(n^2)$, where n is the number of atoms in the two patterns to be compared. However, this test is performed an exponential number of times, thus making the overall computational cost very high. Solutions have been proposed by Nijssen and Kok [27] to gain better performances in the general case of relational association rules. In the context of multiple-level relational association rules, different strategies have been identified by Lisi and Malerba [20]. The **candidate evaluation** phase is performed by comparing the support of the candidate pattern with the minimum support threshold set for the level being explored. If the pattern turns out not to be a large one, it is rejected. As for the support count, the candidate is transformed into an existential query whose answer set supplies all the substitutions that make the pattern true in $D(S)$. In particular, the number of different bindings for the variable which is the placeholder for reference objects is assumed as the absolute frequency of the pattern in $D(S)$.

A rough preliminary remark on the computational complexity of SPADA leads to the notorious trade-off between expressiveness and efficiency in first-order representations. Indeed, it is well known that a simple matching of two expressions with commutative and associative operators (such as the logical OR of atoms in a clause) is NP-complete [12]. Therefore, any known algorithm that checks the coverage of an atom set or that equivalently evaluates a query with respect to a relational database has an exponential complexity. Nevertheless, it has also been proved that queries with up to k atoms, where each atom contains at most j terms, can be evaluated in polynomial time [7]. Whether these constraints are applicable to the domain of spatial data analysis is still under investigation.

Related to efficiency is **scalability**. Indeed, studies on the learnability theory have shown that current ILP algorithms would scale relatively well as the number of examples or facts in the background knowledge increases. However, they would not scale well with the number of arguments of the predicates (relations) involved, and in some cases with the complexity of the patterns being searched. The use of **declarative bias** is usually suggested to improve scalability. It is a set of constraints on spatial patterns and association rules that guide the application of the refinement operator ρ during the candidate generation phase.

Indeed, a refinement step consists of adding one or more atoms from $L[I]$ to the pattern to be refined. The more restrictions we put on the patterns, the smaller the search space, and hence the faster its search. In general, there is a trade-off between the efficiency of an ILP system and the quality of the patterns it comes up with.

2.3. Integrating SPADA with other software components

The application of the ILP approach to spatial databases is made possible by a middle-layer module for feature extraction, as shown in Figure 7.

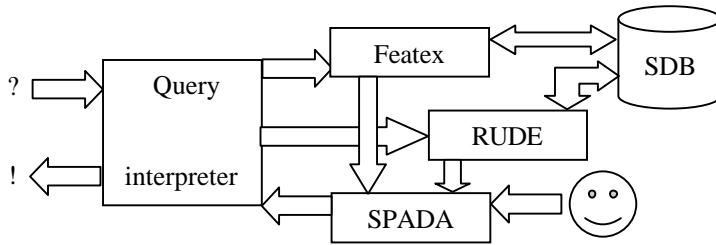


Figure 7: Integration of SPADA with other software modules which support spatial feature extraction (Featex) and discretisation of numerical features (RUDE). Additional input to SPADA, such as declarative bias and background knowledge, is directly provided by the user.

This layer is essential to cope with one of the main issues of spatial data mining, namely the requirement of complex data transformation processes to make spatial relations explicit.

This function is partially supported by the spatial database (SDB), which offers spatial data types in its data model and query language and supports them in its implementation, providing at least spatial indexing and efficient algorithms for spatial join [13]. Thus spatial databases supply an adequate representation of both single objects and spatially related collections of objects. In particular, the abstraction primitives for spatial objects are point, line and region. Among the operations defined on spatial objects, spatial relationships are the most important because they make it possible, for example, to ask for all objects in a given relationship with a query object. Egenhofer and Herring [10] proposed the nine-intersection model to categorise binary topological relations between arbitrary spatial objects. Examples are the relation **meet** between two regions and the relation **crosses** between a region and a line. The nine-intersection model is implemented in the Oracle Spatial cartridge to support the computation of some topological relations.

Many spatial features (relations and attributes) can be extracted from spatial objects stored in SDB. They can be categorised as follows:

- (i) geometric, that is, based on the principles of Euclidean geometry;
- (ii) directional, that is, regarding relative spatial orientation in two or three dimensions;
- (iii) topological, that is, binary relations that preserve themselves under topological

transformations such as translation, rotation and scaling;

- (iv) hybrid, that is, features which merge properties of two or more of the previous three categories.

This variety requires for the development of a feature extractor module, named Featex, which also enables the coupling of SPADA with the SDB. Featex is implemented as an Oracle package of procedures and functions implemented in the PL-SQL language. In this way, it is possible to formulate complex SQL queries involving both spatial and aspatial data (e.g. census data). The set of spatial features that can be extracted by this module is reported in Table 5.

Since SPADA, like many other association rule mining algorithms, cannot process numerical data properly, it is necessary to perform a discretisation of numerical features with a relatively large domain. For this purpose, we have implemented the relative unsupervised discretisation algorithm RUDE [21] which proves to be suitable for dealing with numerical data in the context of association rule mining. At the end of all this data processing, query results are stored in temporary database tables. An ad hoc PL-SQL function transforms these tuples into ground Datalog facts of $D(S)$.

Table 5: Spatial features extracted by the feature extractor module

Feature	Meaning	Type	Values
almost_parallel(Y,Z)	Parallelism relation between Y and Z	Hybrid relation	{true, false}
almost_perpendicular(Y,Z)	Perpendicularity relation between Y and Z	Hybrid relation	{true, false}
density(Y,Z)	Area(Y)/Area(Z)	Hybrid relation	Real
direction(Y)	Geographic direction of object Y	Directional attribute	{north, east, north_west, north_east}
distance(Y,Z)	Distance between Y and Z	Geometrical relation	Real
layer_name(Y)	Object Y type	Aspatial attribute	Layer name
line_shape(Y)	Object Y shape	Geometrical attribute	{Straight, curvilinear}
relate(Y,Z)	Topological relation between Y and Z	Topological attribute	Type of topological relation

3. Application to Stockport census data

In the context of the SPIN! project we investigated the application of spatial data mining techniques to some issues reported in the unitary development plans (UDP) of Stockport, one of the 10 metropolitan districts of Greater Manchester, United Kingdom.

3.1. The data

Spatial analysis is made possible by the use of the Ordnance Survey's digital maps of the district, where several interesting layers are available, namely ED/ward/district boundaries, roads, bus priority lanes, and so on. In particular, Stockport is divided into 22 wards for a total of 589 EDs. By joining UK 1991 census data available at the ED summarisation level with ED spatial objects, it is possible to investigate socioeconomic issues from a spatial viewpoint. In total 89 tables, each having 120 attributes on average, have been made available for policy analysis. Census attributes provide statistics on the population (resident and present at the census time, ethnic group, age, marital status, economic position, and so on), on the households in each ED (number of households with n children, number of households with n economically inactive people, number of households with two cars, and so on) as well as on some services available in each ED (e.g. number of schools).

For the application of our spatial association rule mining method we have focused our attention on transportation planning, which is one of the key issues in UDP. In the following subsection, we report results for the problem of characterising the area crossed by the M63 motorway. For another application to the accessibility of the area around the Stepping Hill Hospital, see the paper by Malerba et al. [23].

3.2. Characterising the area crossed by the M63 motorway

One of the problems is a decision-making process concerning the M63 motorway. More precisely, we are asked to describe the area of Stockport served by the M63 (i.e. the wards of Brinnington, Cheadle, Edgeley, Heaton Mersey, South Reddish) from the sociological viewpoint, in order to provide some hints for transport planners. The data considered in this analysis concerns census statistics on commuters. The description of the area is expressed by some spatial association rules at two levels of granularity. A hierarchy for the Stockport ED layer has been obtained by grouping EDs on the basis of the ward they belong to (see Figure 8) and expressed as Datalog facts in BK.

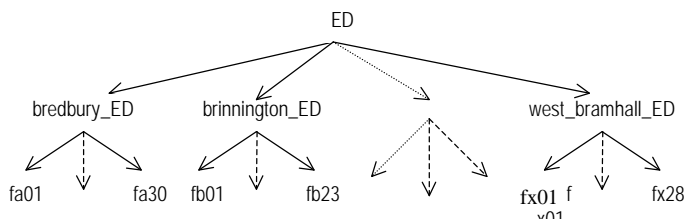


Figure 8: An is_a hierarchy for the Stockport ED layer

Spatial association rules should relate EDs crossed by the M63 (reference objects) to EDs in the area served by the M63 (task-relevant objects) (see Figure 9).

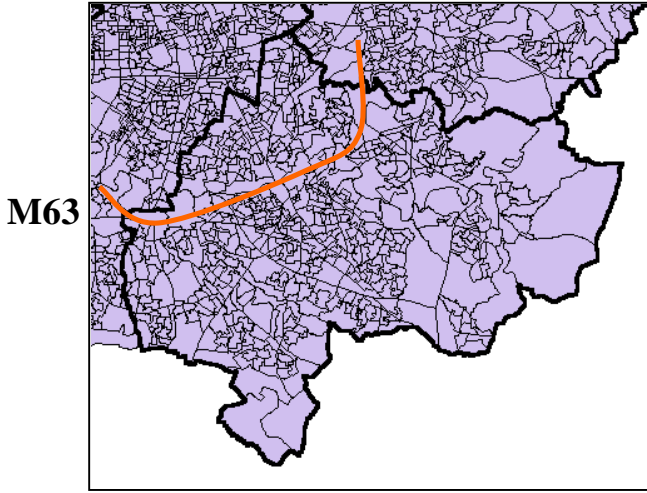


Figure 9: Stockport district and its EDs crossed by the M63 motorway

The relations of intersection (EDs–motorways) and adjacency (EDs–EDs) have been extracted for the area of interest and transformed into Datalog facts of $D(S)$. The following census attributes have been selected for this experiment:

- s820161, persons who work outside the district of residence and drive to work;
- s820213, employees and self-employed workers who reside in households with three or more cars and drive to work;
- s820221, employees and self-employed workers who reside in households with three or more cars and work outside the district of residence.

Since they refer to residents aged 16 and over, they have been normalised with respect to the total number of residents aged 16 and over (s820001). Moreover, they have been discretised by RUDE, since they are all numeric (more precisely, integer valued). At the end of this transformation process, each ED is described by three ground atoms in $D(S)$, namely $dr_out(X, [a..b])$, $cars3_dr(X, [a..b])$, $cars3_out(X, [a..b])$, where X denotes an ED, while $[a..b]$ is one of the intervals returned by RUDE.

The key atom defining the reference objects in S is $ed_on_M63(X)$, which is intensionally defined in the BK by means of the following rule:

$ed_on_M63(X) :- intersect(X, m63)$

The BK also includes the declarative specification of some rules for spatial qualitative reasoning, namely

$\text{can_reach}(X, Y) \text{---} \text{intersect}(X, \text{m63}), \text{intersect}(Y, \text{m63}), Y \neq X.$

$\text{close_to}(X, Y) \text{---} \text{adjacent_to}(X, Z), \text{adjacent_to}(Z, Y), Y \neq X.$

Finally, the following thresholds for support and confidence were defined: $\text{min_sup}[1] = 0.7$ and $\text{min_conf}[1] = 0.9$ at the first level, and $\text{min_sup}[2] = 0.5$ and $\text{min_conf}[2] = 0.8$ at the second level.

SPADA was run on the $D(S)$ obtained. The runtime was 331 seconds for association rules at granularity level 1, and 310 seconds for level 2 (data refers to a Pentium III 1 GHz PC with 256 Mb RAM).

Initially, the system returned 12 925 frequent patterns out of 74 338 candidate patterns, for a total of 12 466 strong rules. By analysing them we observed that some were actually useless, since they did not relate spatial data to census data. In other words, some association rules were pure spatial patterns, such as the following:

$\text{ed_on_M63}(X), \text{can_reach}(X, Y) \ni \text{is_a}(Y, \text{ward_on_m63_ED}) \quad (90.0 \%, 100.0 \%),$

which states that if an ED (Y) in the area served by the M63 can be reached from an ED crossed by the M63, then that ED is certainly (100 % confidence) an ED of a ward crossed by the M63. Despite the high support and confidence, this pure spatial pattern is of no interest for transport planners.

In a second run, we decided to constrain the search to patterns containing at least one of the census attributes $\text{dr_out}(X, [a..b])$, $\text{cars3_dr}(X, [a..b])$ and $\text{cars3_out}(X, [a..b])$. This is possible by specifying the following declarative bias:

$\text{pattern_constraint}([\text{dr_out}(_, _), \text{cars3_dr}(_, _), \text{cars3_out}(_, _)], 1)$

where the first argument of the predicate *pattern_constraint* is the list of atoms to include in the relational pattern, while the second argument is the minimum number of required atoms of the list.

The system generated 10 513 strong association rules in 1 520 seconds (time increased because of constraint checking for each generated pattern). Some of them have a very high support and confidence and provide the expert with some hints on the habits of commuters, such as the following association rule discovered at level 2:

$\text{ed_on_M63}(X), \text{close_to}(X, Y), \text{is_a}(Y, \text{Bedgeley_ED}) \ni$
 $\text{cars3_out}(X, [0.0..0.037]), \text{cars3_dr}(X, [0.0..0.037]) \quad (100 \%, 100 \%),$

which states that ‘if an ED crossed by the M63 (X) is close to another ED of the ward of Bedgeley (Y), then in that ED the percentage of people living in households with three or more cars and driving out of the district to work is very low (less than 4 %)’. It is important to point out that this is simply an association and does not define any kind of cause–effect relationship between the place where people live and their social habits. Another interesting

spatial association rule at the same granularity level is the following:

$$\text{ed_on_M63}(X), \text{can_reach}(X,Y) \rightarrow \text{is_a}(Y, \text{heaton_mersey_ED}), \\ \text{dr_out}(Y, [0.2857..0.4782]), \text{cars3_out}(Y, [0.0..0.037]) \quad (80.0 \%, 88.88 \%),$$

which states that ‘if an ED Y in the M63 area can be reached from another one crossed by the M63 motorway (X), then it is in the Heaton Mersey ward and has quite a high percentage of people that drive to work but do not live in households with three or more cars’.

Finally, we decided to constrain the search space further, by asking only for those spatial patterns involving EDs where people have the same commuting habits. This time the first argument of the predicate **pattern_constraint** is a list of sub-lists, where each sub-list denotes a conjunction of atoms to be included in the relational patterns. In particular, we have defined the following declarative bias:

$$\text{pattern_constraint}([[\text{dr_out}(X,Z), \text{dr_out}(Y,Z), X \neq Y], \\ [\text{cars3_dr}(X,Z), \text{cars3_dr}(Y,Z), X \neq Y], [\text{cars3_out}(X,Z), \text{cars3_out}(Y,Z), X \neq Y]], 1).$$

SPADA found only 345 strong rules (79 for level 1 and 266 for level 2) in about 833 seconds. The following is an example of association found by the system at the granularity level:

$$\text{ed_on_M63}(A) \rightarrow \text{can_reach}(A,B), \text{is_a}(B, \text{cheadle_ED}), \text{can_reach}(A,C), C \neq B, \\ \text{is_a}(C, \text{edgeley_ED}), \text{cars3_dr}(C, [0.0..0.037]), \text{cars3_dr}(B, [0.0..0.037]) \\ (90 \%, 90 \%),$$

which states that from an ED crossed by the M63 it is possible to reach (by the same motorway) two EDs, one in Cheadle and one in Edgley, with the same low percentage of people living in families with three or more cars and driving out of the district to work.

4. Conclusions

In the above application, we have seen that some of the discovered rules actually convey new knowledge. However, the search for these ‘nuggets’ requires a lot of tuning and efforts by the data analyst in order to constrain the search space properly and discard most of the obvious or totally useless patterns hidden in the data. This is typical of exploratory data analysis and SPADA can be considered one of the most advanced tools that data analysts currently use in their iterative knowledge discovery process.

One of the main limitations of SPADA, which is also a problem of many other relational data mining algorithms, is the requirement of some expertise in data and knowledge engineering. Indeed, the user should know how data are organised in the spatial database (e.g. layers and physical representation of objects), the semantics of spatial relations that can be extracted from digital maps, the meaning of some parameters used in the

discretisation process and in the generation of spatial association rules, as well as the correct and most efficient way to specify the domain knowledge and declarative bias. In future work, we will investigate some ‘interestingness measures’ of rules for presentation purposes, so that the user can browse the output XML file of spatial association rules as simply as possible. In addition, we intend to study the relation with ‘symbolic data analysis’ [3] and the possibility of using the software developed in the context of the SODAS project for the analysis, summarisation and visualisation of rules obtained by generalising spatial objects covered by some spatial association rules returned by SPADA.

5. Acknowledgements

The authors thank Jim Petch, Keith Cole and Mohammed Islam (MIMAS, University of Manchester, England) and Chrissie Gibson (Department of Environmental and Geographical Sciences, Manchester Metropolitan University, England) for providing access to census data and digital OS maps of Stockport Manchester. The work presented in this paper is in partial fulfilment of the research objectives set by the IST European project SPIN! (**S**patial mining for data of public **i**nterest) and by the MURST COFIN-2001 project ‘Methods for the extraction, validation and representation of statistical information in a decision context’. Thanks to Lynn Rudd for her help in reading the paper.

6. References

- [1] Agrawal, R. and Srikant, R., ‘Fast algorithms for mining association rules’, *Proceedings of the 20th VLDB conference*, Santiago, Chile, 1994.
- [2] Bergadano, F. and Gunetti, D., *Inductive logic programming: from machine learning to software engineering*, The MIT Press, Cambridge, MA, 1996.
- [3] Bock, H. H. and Diday, E. (eds.), *Analysis of symbolic data — Exploratory methods for extracting statistical information from complex data*, Studies in classification, data analysis, and knowledge organisation series, Vol. 15, Springer-Verlag, Berlin, 2000.
- [4] Ceri, S., Gottlob, G. and Tanca, L., ‘What you always wanted to know about Datalog (and never dared to ask)’, *IEEE transactions on knowledge and data engineering*, Vol. 1, No 1, 1989, pp. 146–166.
- [5] Dehaspe, L. and De Raedt, L., ‘Mining association rules in multiple relations’, Lavrac, N and Dzeroski, S. (eds), *Inductive logic programming*, LNCS 1297, Springer-Verlag, Berlin, 1997, pp. 125–132.
- [6] Dehaspe, L. and Toivonen, H., ‘Discovery of frequent Datalog patterns’, *Data mining and knowledge discovery*, Vol. 3, No 1, 1999, pp. 7–36.

- [7] De Raedt L. and Dzeroski, S., 'First order jk-clausal theories are PAC-learnable' *Artificial Intelligence*, Vol. 70, 1994, pp. 375–392.
- [8] De Raedt, L., *Interactive theory revision*, Academic Press, London, 1992.
- [9] Dzeroski, S. and Lavrac, N. (eds), *Relational data mining*, Springer-Verlag, Berlin, 2001.
- [10] Egenhofer, M. J. and Herring, J. R., 'Categorising binary topological relations between regions, lines, and points in geographic databases', Egenhofer, M. J., Mark, D. M. and Herring, J. R. (eds.), *The nine intersection: formalism and its use for natural-language spatial predicates*, 1994, pp. 183–271.
- [11] Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., 'From data mining to knowledge discovery: an overview', Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds), *Advances in knowledge discovery in databases*, AAAI Press/The MIT Press, 1996, pp. 1–34.
- [12] Garey, M. R. and Johnson, D. S., *Computers and intractability*, W. H. Freeman and Co., San Francisco, California, 1979.
- [13] Güting, R. H., 'An introduction to spatial database systems', *VLDB Journal*, Vol. 3, No 4, 1994, pp. 357–399.
- [14] Han, J. and Fu, Y., 'Discovery of multiple-level association rules from large databases', Dayal, U., Gray, P. M. D. and Nishio, S. (eds), *VLDB'95 — Proceedings of the 21st international conference on very large databases*, Morgan-Kaufmann, 1995, pp. 420–431.
- [15] Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Stefanovic, N., Xia, B. and Zajane, O. R., 'DBMiner: a system for mining knowledge in large relational databases', *Proceedings of the 1996 international conference on data mining and knowledge discovery (KDD'96)*, Portland, Oregon, 1996, pp. 250–255.
- [16] Han, J., Koperski, K., Stefanovic, N., 'GeoMiner: a system prototype for spatial data mining', Peckham, J. (ed.), *Sigmod 1997 — Proceedings of the ACM–Sigmod international conference on management of data*, Sigmod, Record 26, No 2, 1997, pp. 553–556.
- [17] Helft, N., 'Inductive generalisation: a logical framework', Bratko, I. and Lavrac, N. (eds), *Progress in machine learning*, Sigma Press, 1987, pp. 149–157.

- [18] Koperski, K., Adhikary, J. and Han, J., ‘Spatial data mining: progress and challenges’, *Proceedings of the workshop on research issues on data mining and knowledge discovery*, Montreal, Canada, 1996.
- [19] Lavrac, N. and Dzeroski, S., *Inductive logic programming: techniques and applications*, Ellis Horwood, Chichester, 1994.
- [20] Lisi, F. and Malerba, D., ‘Efficient discovery of multiple-level patterns’, *Atti del Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati SEBD 2002*, 2002, pp. 237–250.
- [21] Ludl, M.-C. and Widmer, G., ‘Relative unsupervised discretisation for association rule mining’, Zighed, D. A., Komorowski, H. J. and Zytkow, J. M. (eds), *Principles of data mining and knowledge discovery*, LNCS 1910, Springer-Verlag, 2000, pp. 148–158.
- [22] Malerba, D., Esposito, F., Lanza, A. and Lisi, F. A., ‘Machine learning for information extraction from topographic maps’, Miller, H. J. and Han, J. (eds), *Geographic data mining and knowledge discovery*, Taylor and Francis, London, 2001, pp. 291–314.
- [23] Malerba, D., Lisi, F. A., Appice, A. and Sblendorio, F., ‘Mining spatial association rules in census data: a relational approach’, *Proceedings of the ECML/PKDD’02 workshop on mining official data*, University Printing House, Helsinki, 2002, pp. 80–93.
- [24] Mannila, H. and Toivonen, H., Levelwise search and borders of theories in knowledge discovery, *Data mining and knowledge discovery*, Vol. 1, No 3, 1997, pp. 259–289.
- [25] Muggleton, S. (ed), *Inductive logic programming*, Academic Press, London, 1992.
- [26] Nienhuys-Cheng, S.-H. and deWolf, R., *Foundations of inductive logic programming*, Springer, Heidelberg, Germany, 1997.
- [27] Nijssen, S. and Kok, J. N., ‘Faster association rules for multiple relations’, Nebel, B. (ed), *Proceedings of the 17th international joint conference on artificial intelligence*, Morgan Kaufmann, 2001, pp. 891–896.
- [28] Plotkin, G., ‘A note on inductive generalisation’, *Machine intelligence*, No 5, 1970, pp. 153–163.
- [29] Saporta, G., ‘Data mining and official statistics’, *Atti della Quinta Conferenza Nazionale di Statistica*, Rome, 2000, pp. 15–17

Experiences in developing a spatio-temporal information system

Giuseppe Sindoni, Stefano De Francisci, Mario Paolucci and Leonardo Tininini ⁽¹⁾

ISTAT

Via Adolfo Ravà, 150, I-00142 Rome

E-mail: sindoni@istat.it; defranci@istat.it; paolucc@istat.it; tininini@istat.it

Keywords: temporal databases, GIS, data integration, data warehouse

Abstract

The Italian National Statistics Institute is currently integrating its various legacy spatio-temporal data collections. The SIT-IN project has delivered a first release, whose development relied on web and relational technologies to manage data heterogeneity. The final system provides users with many different classes of functions with which to analyse and visualise territorial data. It can be viewed as a spatio-temporal data warehouse, where space and time are the main access dimensions to statistical data, but also where the changes in space over time can be analysed as a preliminary step in the design activity of a statistical survey.

SIT-IN overcomes a drawback of current commercial data warehouse systems, which are unable to cope with the dynamic behaviour of a dimension: that is, its temporal evolution. Another of its features is its ability to realise 'lightweight alliances' with external database tables.

1. Introduction

ISTAT, the Italian National Statistics Institute is currently integrating its various legacy spatio-temporal data collections. The SIT-IN project has delivered a first release [9], whose development relied on web and relational technologies to manage data heterogeneity.

The integrated systems are: a historical database, providing information about the evolution in time of territorial administrative partitions; the institute's GIS, providing the cartography of the Italian territory down to census tract level; a statistical data warehouse, providing spatio-temporal statistical data from a number of different surveys; an address normalising/geo-matching system, providing information about the limits of census tracts.

The final system provides users with many different classes of functions with which territorial data can be analysed and visualised. It can be viewed as a spatio-temporal data warehouse, i.e. a data warehouse where space and time are the main access dimensions to statistical data, but also where space can be analysed according to its temporal mutations as a preliminary step in the design activity of a statistical survey.

(¹) The author was partly supported by CNR-IASI, Viale Manzoni, 30, I-00185 Rome.

The Italian territorial hierarchy is managed and its temporal dynamic behaviour stored in a relational database. This has been designed according to a generalised spatio-temporal model which allows the representation of:

- territorial objects (e.g. regions, towns, census tracts, etc.);
- territorial objects both temporally and spatially (e.g. the polygons representing each temporal version of a territorial object);
- temporal attributes of a territorial object (e.g. name, code or any statistical value which is temporally associated with the object);
- the territorial hierarchy;
- temporal evolution of the territorial hierarchy.

This great flexibility in the management of the space dimension enables the SIT-IN system to overcome a drawback of current commercial data warehouse systems, which are unable to manage the dynamic behaviour of a dimension, i.e. its temporal evolution. Using SIT-IN, the statistical data warehouse can be queried for a given space-time plane or produce a time series of data for a given area of interest, without the need to consider the territorial hierarchy's temporal evolution. The system manages the spatio-temporal relationships in such a way as to present the user with the correct sets of territorial and statistical objects for a given time-stamp or time interval.

Another feature of SIT-IN is its ability to realise 'lightweight alliances' with external database tables. In order to prove the feasibility of a dynamic integration function, a module has been implemented which allows the user to dynamically link a database table containing statistical values referred to a given time-stamp and territory of interest. The linked table can currently be used only to produce a thematic map from its values, but future developments of the system will allow the table's full integration with other statistical data for performance of cross-analysis and other operations.

The paper is organised as follows: Section 2 briefly outlines the systems under integration; in Section 3 the principles underlying the integration process are discussed and in Section 4 the overall architecture of the system is described. Section 5 shows how SIT-IN overcomes some important features lacking in commercial products and illustrates the main functions of the system.

2. The systems under integration

The integrated legacy systems include those listed below.

- Sistat, the territorial history database system, providing information about the temporal evolution of territorial administrative partitions: The system records the administrative

provisions, the date when they become effective and the involved territorial entities (towns, provinces, etc.). Spatial data, however, are not explicitly included.

- Census, the institute's GIS, providing Italian cartography to the census tract level: The layers describing the entities of Italian territorial hierarchy have been 'time-stamped' to provide a cartographic description of the territory evolution. The system is very dynamic both technologically and in its degree of detail. For example, all layers are currently being updated with reference to a collection of fine-grained orthographic photos of the Italian territory of a 1:10 000 scale.
- BDT, a territorial statistical data warehouse, originally designed in the 1980s for mainframe technology, based on a non-relational information system for data dissemination purposes: During the integration process, the information system was migrated to a relational (Oracle) database and a completely new navigational interface was developed, based on the data warehouse paradigm. Aggregate data mainly from the 1971, 1981 and 1991 censuses are available to the town level of detail.
- Sister, an address normalising-geomatching system, providing information about the limits of census tracts (e.g. portions of streets or the sides of town squares).

3. The integration principles

One of the main goals of the SIT-IN project has been to define guidelines for database integration activities involving coexistence with and migration of legacy applications. Our approach to the problem, from the physical data integration point of view, is based on a **loosely coupled federation of databases** [6], [7], [11]: the federated systems continue to exist individually, and are connected through a dynamic data linking mechanism, which somewhat resembles the lightweight alliances approach presented by King et al. [4].

A federated database system is a collection of autonomous cooperating database systems. One of the most important aspects of such an architecture is that the component systems can participate in the federation while concurrently continuing their local everyday operations. Building a federation of databases is essentially a matter of selective and controlled integration of the component systems, which can be managed either by the users of the federation or by its administrator together with the administrator of each single component database. It is worth noting that a federated system may also be a component of another federated system.

The development of a federated database system consists of creating one or more federated schemata upon which query (and eventually updates) are performed. If we have more than one federated schema, we can have a **loosely coupled** federation, where it is the user's responsibility to create and maintain the federation. The system described here is of this type, as will be discussed. Conversely, a **tightly coupled** federated system is where the federation and its administrator(s) are responsible for its creation and maintenance and actively control access to component databases. A tightly coupled system can have only one federated schema.

Our database federation was realised according to the following principles.

Legacy systems are independent from the federation and are fully and totally responsible for the realisation and maintenance of their export scheme, that is, for what they share with the other federated systems. For each overlapping concept, only one component system was chosen as the ‘owner’. For example, the system responsible for data on the administrative history of the Italian territory exports a **snapshot** to the federation and is responsible for its update. The snapshot is used by the federation as the only repository of data about the territorial hierarchy and its time evolution [10]. This concept separation was made possible by an accurate definition of each component’s areas of competence, restricting to small, clearly identifiable parts the areas of interrelation and/or overlapping [2].

Time and space are the main access dimensions for each federated database. For example the territorial data warehouse system is accessed through a set of **navigational metadata**, which allow the user to dynamically define the structure of statistical tables. At run time, the visualised report table refers to the chosen year and set of territorial entities (for example a set of towns).

Data coming from different physical worlds are integrated using relational technologies. For example, the geographical legacy database is run by ESRI Arc/Info, but has been migrated into SDE (Spatial Data Engine), the ESRI spatial indexing system, which allows storage of and efficient access to geographical data through a relational database.

4. SIT-IN architecture

The high complexity of the application context did not permit implementation of a simple client–server system. In fact, the SIT-IN prototype has the three-layer architecture depicted in Figure 1.

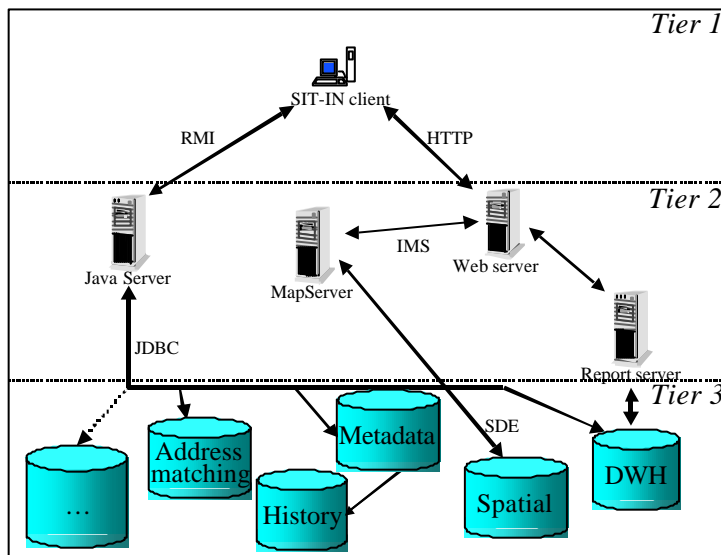


Figure 1: The application deployment architecture.

Tier 3 The data level is composed by a few relational (Oracle) DB instances. On top of the spatial database, the spatial data engine allows efficient access to and manipulation of geographical data. Implementing the above declared principles required a massive usage of mediators [1], [13] to wrap and interconnect the different areas of the systems, based on a collection of **metadata** and **metaqueries** used by the application servers.

One of the main advantages of this layered structure and of the mediator approach is that the number and kind of accessed database instances is not predetermined at start-up time. This allows users to dynamically link their own Oracle data to the system, enabling administrative checking and spatial analysis on this information. As an example of such a mechanism, we implemented a function for dynamic geographic theme building and visualisation on user data. By providing all necessary connection information, the user can link in an external database table, containing a set of statistical values referring to a set of territorial entities, and dynamically generate a thematic map based on such values.

Tier 2 The application layer consists of various application servers, particularly: (i) a home-made Java server handling communications to and from the databases, so separating the user interface from data access; (ii) an **Internet map server** extending the GIS capabilities and allowing cartography serving on the web; (iii) a report server based on ad hoc scripts encapsulating the SAS Intr/Net services and dynamically generating HTML pages of statistical tables.

The use of Java and JDBC allowed easy implementation of the mediator components for each member of the federation. The dynamic integration process was implemented by a meta-querying system [8], [12], part of the SIT-IN metadata management system, a set of database tables containing data about users, linked external tables and values and, above all, data on the administrative hierarchy of the Italian territory. Moreover, the tables

of metadata contain parametric queries, i.e. the parametric SQL scripts required to access territorial data and link them to user data.

Tier 1 System-user interaction is enabled by: (i) a home-made Java applet, which allows flexible multidimensional navigation; (ii) a freeware Java applet, which has been encapsulated to allow the dynamic presentation of geographical/thematic maps; (iii) a ‘classical’ HTTP interaction, based on dynamically generated HTML pages.

To summarise, our system is a loosely coupled federation of database systems, in which one component is the (tightly coupled) federation of the four systems described in Section 2, and the others are dynamically linked user databases.

5. Main functions and solved problems

In this section we illustrate the main functions of the system, with special attention to the implementation of spatio-temporal navigation through aggregate data. We also focus on the main problems of current commercial systems which have been addressed and solved in the context of the SIT-IN project.

5.1. Dynamic dimension instances

Commercial OLAP systems do not consider dynamic dimensions (dimensions which slowly evolve over time, as defined by Mendelzon and Vaisman [5]) such as the spatial partition of a territory. Consider for instance the administrative subdivision of a country territory into a hierarchy of regions, provinces and towns. Several factors affect the dynamics of this hierarchy: a town may be re-assigned to a different province, towns and provinces could be renamed, new towns and provinces can be introduced (typically from part of the territory of other towns and provinces), border ‘adjustments’ could be established by administrative laws. This dynamic behaviour has of course an influence on any time-dependent data analysis process. In SIT-IN, this problem was overcome by implementing a general spatio-temporal data model which completely describes the possible mutations in time of the objects of interest, e.g. object birth and destruction, modification of object properties (borders, name, statistics, etc.) and object inclusion in a less detailed object. Our model can be viewed as an extension of ‘the description of change with respect to states of existence and non-existence for identifiable objects’ proposed by Hornsby and Egenhofer [3]. The basic idea here is to associate each territorial object with a unique identifier and a time interval representing the object’s range of validity. In our model the properties and relationships between objects have also been time-stamped, particularly the property of being child/parent of other objects, names, boundaries, etc. Finally, a directed stratified graph, called **transfer graph**, is defined among territorial objects to represent the transfer of a portion of territory from one object to another.

For example, Figure 2 schematically represents the evolution of an Italian region (Lombardia) after the creation of two new provinces in 1992: the new-born province of Lodi comprises some

territory from the province of Milano and the new-born province of Lecco comprises territory taken from both Como and Bergamo.

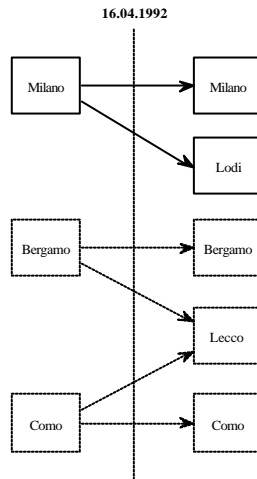


Figure 2: Temporal evolution of Lombardia provinces

This schema can also be considered as an implementation of a multidimensional model with space acting as a ‘dynamic dimension’ [5]. Figure 3 shows the user interface for the definition of a spatio-temporal selection.

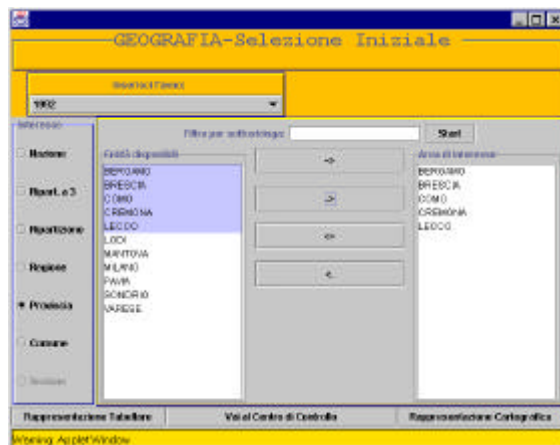


Figure 3: Spatio-temporal selection

The choice of time-stamp implicitly defines not only the proper set of territorial objects for each administrative level, i.e. the objects which effectively existed at that time, with their corresponding property values (name and code), but also the inclusion hierarchies valid at that time. The result of the drill-down from region to province level consequently depends on the chosen time-stamp. For

example, Lombardia drill-down to the province level lists nine provinces for 1991, but gives 11 provinces in 1993 (including the new provinces of Lecco and Lodi, see Figure 3). This selection is the basis for any further multidimensional navigation through aggregate data.

5.2. Spatio-temporal visualisation of data and territory evolution

Existing commercial GISs and DBMSs still do not have enough features to effectively visualise the temporal aspects of spatial data. In fact, GISs structure data in layers, i.e. sets of spatial objects sharing the same property, without providing tools to model their temporal evolution; while DBMSs, although offering specific features for the management of temporal and spatial data, lack primitives and operators for the effective integrated management of the two aspects. In SIT-IN, layers are dynamically generated according to the chosen time-stamp, and data referring to territorial objects in a given time are mapped to the appropriate polygon, thus enabling complex integrations such as the construction of thematic maps based on user data (see below).

The temporal evolution of the area of interest is dynamically visualised on a set of temporal geographic maps. Here, each map is dynamically built with reference to the correct set of spatial objects, which are retrieved by a spatio-temporal query. For each distinct **version** of the territorial partition a different layer is generated and object evolution highlighted by establishing a 1:1 correspondence between objects and colours (see Figure 4.) The system determines the ‘relevant’ objects to display by starting from the user’s selection and then calculating the connected components on the evolution graph.

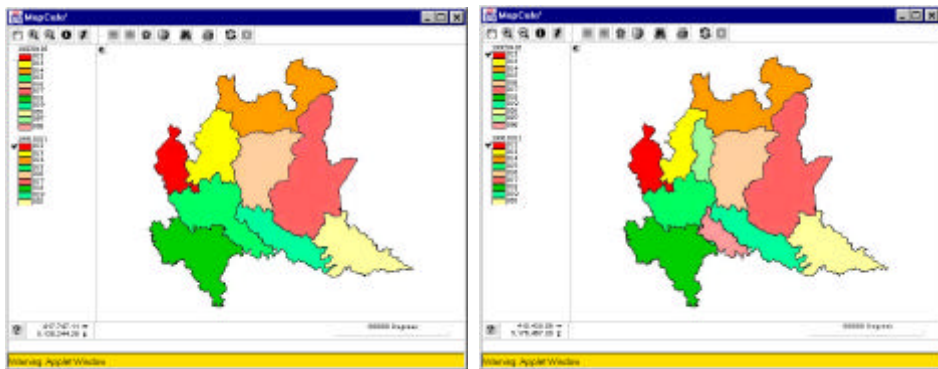
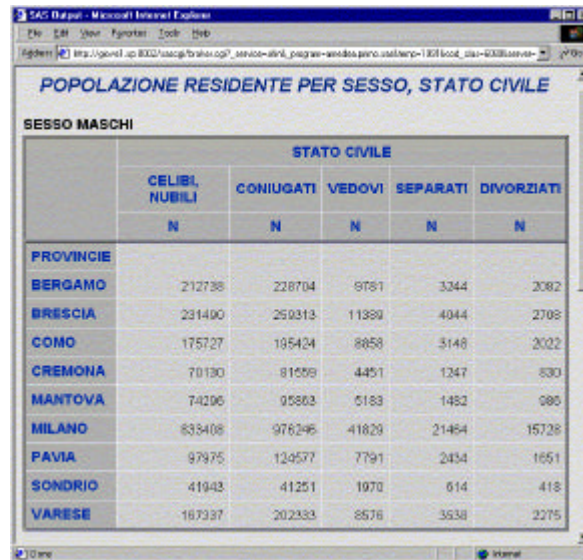


Figure 4: Visualisation of the result of a historical query: maps of the Lombardia region in 1991 and 1992

5.3. Multidimensional navigation and time series

Spatio-temporal selection is used as the access plane to navigate aggregate data. Each dynamic report table refers to the chosen time-stamp and set of territorial entities (Figure 5). Dynamic dimensions are effectively managed: the system is indeed able to restrict navigation to the proper dimension instance, i.e. the instance relative to the chosen time-stamp. For example, for a given 1991 survey, only classification variables and hierarchies which were observed in that survey can

be selected, rolled-up or drilled down. Conversely, by first selecting one or more classification variables, only the ‘facts’ that were observed and classified accordingly can be selected. In this way the user always has the maximum freedom compatible with the selections already performed and is led to express a ‘successful’ query, i.e. a query corresponding to data actually in existence.



POPOLAZIONE RESIDENTE PER SESSO, STATO CIVILE					
SESSO MASCHI					
	STATO CIVILE				
	CELIBI, NUBILI	CONIUGATI	VEDOV	SEPARATI	DIVORZIATI
	N	N	N	N	N
PROVINCIE					
BERGAMO	212738	228704	8781	3244	2082
BRESCIA	231400	259313	11389	4944	2708
COMO	175727	195424	8958	3148	2022
CREMONA	70130	91559	4451	1247	830
MANTOVA	74296	95863	5183	1482	985
MILANO	833408	976246	41829	21464	15728
PAVIA	97975	124577	7791	2434	1651
SONDRIO	41943	41251	1970	614	418
VARESE	167337	202333	8576	3538	2275

Figure 5: A multidimensional table: 1991 resident population in Lombardia, males classified by marital status

A time-slice selection of spatial aggregated data can be ‘extruded’ along the time dimension to produce a time series of spatial data which takes into account the heterogeneity of the space dimension. In particular, for each year, the correct composition of the Lombardia region is considered and by exploiting historical information made available by the system, the researcher can perform effective comparisons along the time dimension (see Figure 6).

	anno					
	1988	1989	1990	1991	1992	1993
	N	N	N	N	N	N
PROVINCIE						
BERGAMO	820228	824804	831885	832370	817264	824166
BRESCIA	1036112	1030548	1045419	1044699	1050405	1055881
COMO	787942	790788	795728	795756	525102	528282
CREMONA	327846	327536	328027	327764	328867	329885
LECCO					298274	299795
LODI					185553	187273
MANTOVA	370892	370980	370832	369314	369410	369190
MILANO	3685433	3686636	3692204	3620626	3740806	3734206
PAVIA	498054	496753	496040	490478	490618	491988
SONDRIO	176167	176485	178769	175453	176015	176371
VARESE	796267	798782	802524	796981	800291	803986

Figure 6: A time series on a dynamic dimension: yearly resident population in Lombardia from 1988 to 1993

Two further features are under development to improve comparability of evolving territories.

- The first is based on the transfer graph and the calculation of the so-called **minimum invariant territories**. For a given time interval, a minimum invariant territory is such that: (i) it always comprises a collection of entire territorial entities in the considered time interval; (ii) property (i) does not hold for any strict subterritory. In the example of the Lombardia provinces from 1991 to 1993, the union of Bergamo and Como provinces in 1991 (or, equivalently, the union of Bergamo, Como and Lecco provinces in 1993) produces a minimum invariant territory. In essence, it is a set of territorial objects whose union does not change in the given time interval. In the (rather frequent) case that the aggregate function is summable, the aggregates corresponding to the minimal invariant territories are obtained from the connected component of the transfer graph and then by summing the correspondent values. So, in our example we can only compare the sum of the values for Bergamo and Como in 1999 with the sum of the values for Como, Bergamo and Lecco after the territorial change (which happened on 16 April 1992).
- The second technique is based on **flattening** the territory evolution. The idea is to recalculate the aggregates directly from microdata (or highly detailed aggregates such as those at census tract level), by mapping the data to the territorial partition at a given time (usually different from the partition in effect when the microdata was collected). For example, in the case of census 1991 data, the Lecco province population in 1991 can be calculated (even though the province itself was established in 1992) by mapping the census microdata to the 1993 territorial partition and then aggregating them accordingly. In practice, the problems related to territory evolution are solved by fixing a precise instant of time and recalculating the aggregates so that they all refer to the same territorial partition, namely the one corresponding to the chosen time-stamp. Unlike the first technique, it is not always possible to apply the

second as it requires access to micro (or highly detailed aggregate) data, which is not always available. Calculations in this case can be only approximate and result in estimated values.

6. Dynamic integration of user data

Users can dynamically link their own data to the system. As an example of such a mechanism, we implemented a function for dynamic geographic theme building and visualisation based on user data. The function is based on the metaquerying technique: the database link to the user's table is created from the parametric statement stored in the metadatabase, using the connection parameters and table information provided by the user. The system automatically validates the territorial codes in the linked table by comparing them with those obtained from the historical database (again by executing the instance of a parametric query). Finally, the specified data is joined to the set of polygons corresponding to the specified time and territorial hierarchy and the thematic map is generated and displayed to the user.

Figure 7 shows the dynamic thematic map generation obtained from the population density values of the Lombardia provinces in 1991 and 1994. Note that the values are correctly mapped to distinct polygons in the two years.

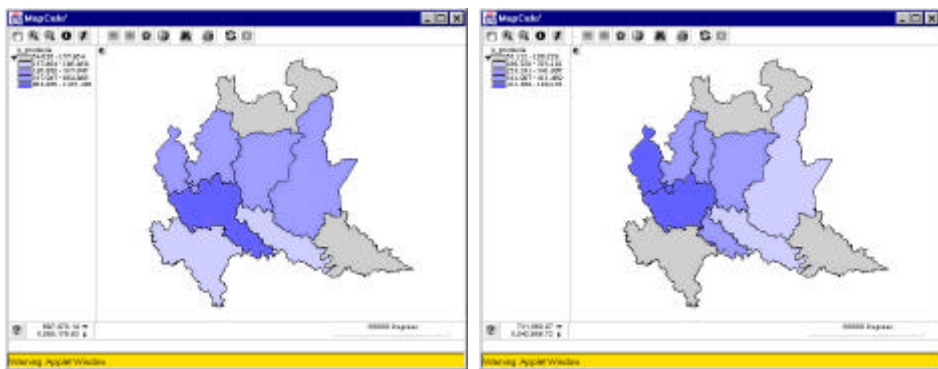


Figure 7: Thematic maps: population densities of the Lombardia provinces in 1991 and 1994

7. Acknowledgements

The authors would like to thank Amedea Ambrosetti, Cristina Bedeschi, Orietta Gargano, Rossella Molinaro, Paola Patteri and Pina Ticca for their useful comments and invaluable help in the system's development. Thanks also to Marie-Hélène Hayles for help with the English.

8. References

- [1] Critchlow, T., Ganesh, M. and Musick, R., ‘Metadata based mediator generation’, *Proceedings of the third IFCIS conference on cooperative information systems (CoopIS’98)*, 1998, pp. 168–176.
- [2] Fang, D., Hammer, J. and McLeod, D., ‘The identification and resolution of semantic heterogeneity in multidatabase systems’, *Proceedings of international workshop on interoperability in multidatabase systems*, Kyoto, April 1991.
- [3] Hornsby, K. and Egenhofer, M. J., ‘Identity-based change: A foundation for spatio-temporal knowledge representation’, *International Journal of Geographical Information Science*, Vol. 14, No 3, 2000, pp. 207–224.
- [4] King, R., Novak, M., Och, C. and Vélez, F., *Sybil: Supporting heterogeneous database interoperability with lightweight alliance*, NGITS, 1997.
- [5] Mendelzon, A. O. and. Vaisman, A. A., ‘Temporal queries in OLAP’, *International conference on very large data bases (VLDB’00)*, Cairo, Egypt, 10–14 September 2000, pp. 242–253.
- [6] Naumann, F., Leser, U. and Freytag, J. C., ‘Quality-driven integration of heterogeneous information systems’, technical report, *Informatik Bericht 117*, Humboldt University, 1999.
- [7] Navathe, S. B. and Donahoo, M. J., ‘Towards intelligent integration of heterogeneous information sources’, *Proceedings of the sixth international workshop on database re-engineering and interoperability*.
- [8] Neven, F., Van den Bussche, J., Van Gucht, D. and Vossen, G., ‘Typed query languages for databases containing queries’, *Information Systems*, Vol. 24, No 7, 1999, pp. 569–595.
- [9] Paolucci, M., Sindoni, G., De Francisci, S. and Tininini, L. ‘Sit-in on heterogeneous data with Java, http and relations’, *Workshop on Java and databases: persistent options*, in conjunction with NetObject.Days conference, 2000.
- [10] Pissinou, N., Snodgrass, R. T., Elmasri, R., Mumick, I. S., Tamer Özsu, M., Pernici, B., Segev, A., Theodoulidis, B. and Dayal U., ‘Towards an infrastructure for temporal databases’, *Sigmod Record*, Vol. 23, No 1, March 1994, pp. 35–51.
- [11] Sheth, A. P. and Larson, J. A., ‘Federated database systems for managing distributed, heterogeneous, and autonomous databases’, *ACM computing surveys*, Vol. 22, No 3, 1990, pp. 183–23.

- [12] Van den Bussche, J., Van Gucht, D. and Vossen, G., 'Reflective programming in the relational algebra', *Journal of Computer and System Sciences*, Vol. 52, No 3, June 1996, pp. 537–549.
- [13] Wiederhold, G., 'Mediators in the architecture of future information systems', *IEEE Computer*, Vol. 25, No 3, 1992, pp. 38–49.

FORUM

This section of the ROS Journal contains contributions, which are mostly for information purposes. Such contributions should present reports on:

- specific statistical research projects and programmes;
- statistical research activities in official statistical institutes;
- experience on practical application of new techniques and technologies for statistics;
- experience on transfer of technologies and know-how both from the perspectives of those making the transfer and those to whom the transfer is being made;
- book reviews, etc.;
- other information of general interest.

Imperatively, papers published in the section have not been put through the usual full review process. Their review has been light and has been dictated by the nature of the paper.

Towards standardisation of survey outcome categories and response rate calculations

Peter Lynn (*), Roeland Beerten (**), Johanna Laiho (***) and Jean Martin (**)

(*) *Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom; e-mail: p.lynn@essex.ac.uk*

(**) *Office for National Statistics, 1 Drummond Gate, London SW1V 2QQ, United Kingdom; e-mail roeland.beerten@ons.gov.uk and jean.martin@ons.gov.uk*

(***) *Statistical Methodology R & D, Statistics Finland, PO Box 5V, FIN-00022 Tilastokeskus; e-mail: johanna.laiho@stat.fi*

Keywords: Contact rates; cooperation rates; eligibility; field procedures; refusal rates; response rate definitions

Abstract

Survey response rates are important process quality indicators and are used for many purposes. However, attempts to compare response rates — across surveys, years, organisations and countries — are severely hampered by inconsistencies in the use of survey outcome categories and in the calculation of response rates based upon these categories. In this article, we highlight some of the main issues. With regard to outcome categories, these include the structure of the coding schema, the definitions of the categories and field implementation. With regard to response rate calculation, main issues include the suitability of different rates for different purposes, how to treat uncertainty regarding eligibility of sample units and whether or not to weight the data. We illustrate some of these issues using data from UK surveys and we suggest the possible form of some solutions.

1. Introduction

Response rates are generally considered to be amongst the most important indicators of survey process quality for surveys of households. They are widely reported and quoted in survey technical reports, and survey commissioners often specify a response rate target as an indicator of the quality they wish to achieve. Response rates are frequently used to compare survey quality between surveys, survey organisations and countries and over time. Response rates are important as non-response can introduce bias. Methodological research (e.g., [7], [8], [9]) shows that non-respondents generally differ from respondents in important characteristics. In the presence of bias, high precision does not guarantee accurate estimation. A second reason for the importance of response rates is that low response means that fewer cases are available for analysis, thus reducing the precision of estimates. Given the importance attached to this survey process quality indicator it is surprising to observe a lack of standardisation in response rate calculations and the outcome categories on which they are based. This will inevitably lead to invalid or misleading comparisons.

Differences in survey design may affect comparisons of response rates and outcome categories between surveys. Different design features will lead to the use of different

outcome categories and different response rate calculations. However, apart from survey design differences, there are differences that could be avoided in the practice of assigning outcome categories and calculating response rates. Most organisations have their own procedures to define survey response rates, often developed over time and with differences between different surveys. Differences in response rates caused by differences in organisational practices might be reduced or avoided if there were generally agreed standards in this area.

There are not only problems in comparing response rates between surveys or organisations. Increasingly surveys are being coordinated or commissioned at international level (e.g. European). Cross-national and international comparisons of survey data are set to become even more important because of an increasing need for policies at this level. There is evidence that the lack of standards for response rate calculation leads to invalid comparisons of survey quality between countries. In a study by de Heer [3] for example, response rates on the labour force survey are compared between 16 countries, but the study admits there is a lack of precise definitions and formulae to calculate the response rates. A report on the methodology of the pilots of the Eurostat time use survey also highlights the lack of information regarding response rate definitions and the difficulties in making cross-national comparisons [16]. On the European household panel survey (ECHP) attempts were made to standardise definitions of the main outcome categories and response rate calculations. However, other differences in survey implementation meant that comparison of rates between countries were not always valid. It is proposed that the European social survey will introduce standardised contact and outcome definitions and will undertake centralised non-response analysis in order to maximise comparability as well as publishing metadata regarding sampling and data collection procedures to aid interpretation [4].

2. History and context

Although the absence of standard definitions of outcome categories and response rates has long been recognised [10], [15], [11], there have been relatively few attempts to introduce standards across survey organisations. A survey of major international and American professional and trade associations involved in survey and market research found that only three out of 14 associations have any kind of guidelines on calculating and reporting response rates [17], [18]. One of the earlier efforts to develop standards in this area was made in the USA by the Council of American Research Organizations (CASRO) in 1982 [6]. More recently, progress has been made by the American Association for Public Opinion Research (AAPOR). Based on the earlier CASRO work, AAPOR published a report with an updated, comprehensive set of outcome codes and operational definitions, including formulae for calculating different response rates [1].

However, the AAPOR recommendations are limited and are not directly applicable to other countries for at least three reasons. First, they deal only with surveys involving a single respondent within a household. However, many surveys collect information from (or regarding) **all** members of the household. Second, they deal only with RDD telephone surveys, in-home surveys based on samples of residential addresses using procedures

common in the USA and mail surveys of specifically named persons. The nature of the sampling methods and sampling frames used for many social surveys in Europe, for example, raises issues that are not dealt with in the AAPOR document. Considerable work is therefore needed before the AAPOR standards can be adapted for use in other countries. This work includes both conceptual development and careful ‘translation’ of existing concepts. Third, the AAPOR document does not provide practical guidance for field implementation, nor deal with a number of technical issues that we feel are important.

In order to develop generic standards that will be applicable across a broad range of countries and survey types, it is important first to recognise the variation — particularly between countries — in constraints and designs. In the case of household surveys, a key dimension of variation concerns the nature of available sampling frames. The main categories are population registers, lists of addresses and area-based approaches. These different frames lead to different sets of possible field outcomes and hence to different outcome categories. The challenge is to develop a generic framework to which these different sets of outcomes can be related in a consistent way.

3. Towards standardisation

Standardisation is required in a number of areas: identifying and defining outcome categories, defining response rates for different purposes and providing metadata to aid the interpretation of these rates.

3.1. Outcome categories

In order for response rates to be comparable, harmonisation has to be achieved at all stages of the survey process, from the early stages of design to final reporting. It is not sufficient merely to use a standard **list** of outcome categories. It is necessary also to employ a standard **definition** of each category. The definitions must contain sufficient detail and must result in a comprehensive and mutually exclusive set of categories. Those applying the categorisation (mainly interviewers and field staff in the case of interview surveys; office staff in the case of postal or web-based surveys) must be adequately trained in the standard interpretation of the definitions. Furthermore, response rates and other outcome rates must be calculated, presented and labelled in a standard way. This requires standard definitions of outcome rates, including specific guidance on the treatment of each outcome category as well as on the labelling and interpretation of the rates.

The key to comparable response rates is the use of standardised final outcome categories. They form the basis for response rate calculations. We believe it is possible to develop standardised final outcome categories that can be adopted — at least to a reasonable level of detail — by all surveys, or all surveys of a particular general type. Surveys will sometimes have specific characteristics that require the use of extra outcome categories in addition to the standard ones, but this must be handled in a way that preserves the standard structure of the outcome category schema. This can be achieved by adopting a hierarchical

structure of outcome categories and permitting the introduction of survey-specific outcome categories as a further level of detail **within** the standard hierarchy.

Each outcome category should have a clear and unambiguous definition, which should be sufficiently detailed to remove reasonable doubt from the choice of appropriate outcome category for almost all situations that are likely to be encountered in practice. The prime aim of this guidance would be to enable choice of the correct (temporary) outcome code on every occasion. An example of such guidance is discussed in Section 5.1 below. Temporary outcome categories could then be processed into final outcome codes with little difficulty. Once standardised outcome categories have been adopted, standard methods for calculating response rates can be introduced. The implementation of standardised outcome categories and the subsequent calculation and presentation of standardised response rates should be tested in pilot surveys.

3.2. Metadata

Appropriate interpretation of response rate comparisons requires awareness of many survey-specific features that can have a big effect on the magnitude of response rates. These features can usefully be considered in two categories. The first is a set of definitions than underpin outcome categories and response rate calculation. These include definitions of the sampling unit, eligibility for the survey, desired respondent and acceptable proxy respondents, complete vs. partial interview/data, partial interview/data vs. unit non-response, etc. The second category of survey-specific features that should influence the interpretation of response rates consists of features of the data collection process such as the nature of the respondent task, constraints upon the fieldwork timing and the volume and nature of contact attempts made.

3.3. Response rates

In addition to the design and implementation features mentioned above, the choice of response rate definition can considerably influence the magnitude of reported response rates. Typically, definitions tend not to be reported or only to be described in rather general terms. There are many possible subtle variations in response rate definition, so fairly precise details of the definition are required in order for interpretations — and particularly comparisons — to be meaningful. Furthermore, there are a number of different rates that can provide useful information for different purposes. Consequently, surveys should be encouraged to publish multiple rates, each following an agreed and useful definition (some recommendations in this area are summarised in Section 4.4 below).

Also, relevant features of the data collection method should be recognised using mode-adapted temporary and final outcome codes. Surveys that involve supplementary data collection instruments (self-completion documents, diaries, physical measurements, blood samples, etc.) in addition to a main interview component will require additional instrument-specific outcome categories and rates, but this should not affect the ability **also** to present overall survey response rates in a standardised way.

4. Some recommendations

4.1. Schema of final outcome categories

We believe that survey outcome categories naturally have a multi-level hierarchical structure, closely related to the hierarchical process of making contact, establishing eligibility and gaining cooperation. Indeed, most outcome categorisations currently in use on major social surveys already have a hierarchical structure (e.g. [1]), though some or all levels of the hierarchy may not be explicit. The first three broad levels of the hierarchical process of establishing outcomes on an interview survey are illustrated in Figure 1. It should be noted that ‘contact’ must be defined. We suggest that it should involve verbal interaction between interviewer and household member (by telephone or in person). In the case of a postal or web survey, the process is in principle very similar, though it is of course much more difficult to identify when and whether ‘contact’ has been made and when a non-response constitutes a refusal or is due to some other reason.

In practice, the stages presented in Figure 1, once all components have been suitably defined, can form the basis of a system of outcome categorisation which covers all possible situations. For interview surveys, the numbered categories may be treated as the first level of a hierarchy of outcome categories, as shown in Figure 2. Within each of these categories, there will be multiple subcategories. An example of a full list of possible standard subcategories for interview surveys of households appears in Appendix A. Figure 3 presents a part of this list and shows the various levels and sub-levels of the category ‘refusals’. Each of the detailed codes can provide information about the nature of the non-response on a particular survey, and will be useful in deciding on measures to reduce non-response. It can be seen that this schema can be applied to a broad range of survey types. For example, category 4.1 applies only to surveys where sample members are given advance notification (e.g. a letter in advance of an interviewer approach). On other surveys, this category simply need not be used. Similarly, 4.2 applies only to address-based samples. On the other hand, extra subcategories can be used on specific surveys, provided the standard structure is maintained. For example, if it is desired to record stated reasons for refusals, these could be made subcategories of 4.3.1, 4.3.2 and/or 4.4. Full definitions of the categories presented in Appendix A can be found in Lynn et al. [13].

We propose that schema should be developed in anticipation of standard adoption at least to the level of detail of the second level of the hierarchy shown in Appendix A. The third level categories might be considered as optional in the sense that not all surveys need record outcome information to this level of detail but those that do should adhere to the standard categorisation. With such a schema, some categories will not apply in some cases, but this should not affect the standard use of the remaining categories. It should be noted that some adaptation to the schema is necessary for postal and other self-completion surveys, in particular to deal differently with the concepts of contact and established eligibility.

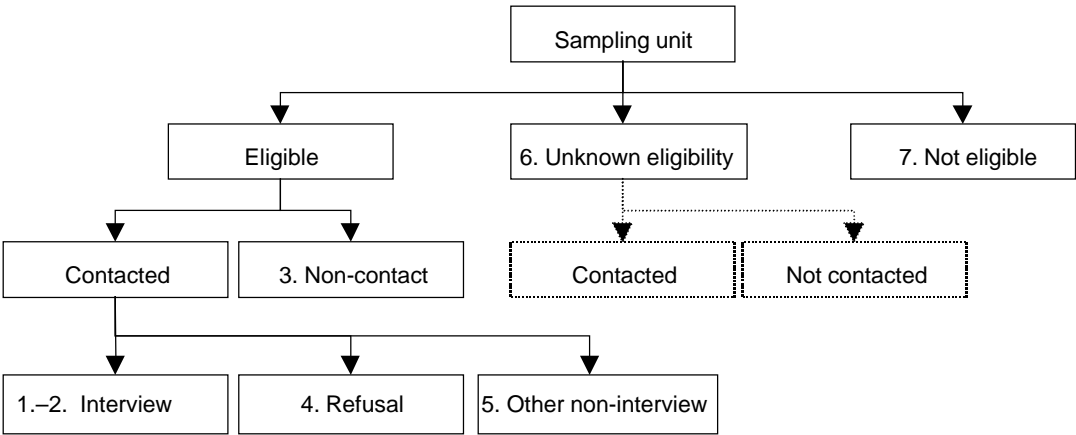


Figure 1: The hierarchical process of establishing survey outcomes

-
- 1. Full interview
 - 2. Partial interview
 - 3. Non-contact
 - 4. Refusal
 - 5. Other non-interview
 - 6. Unknown eligibility
 - 7. Not eligible
-

Figure 2: The first level of a hierarchy of outcome categories ⁽¹⁾

4. Refusal	4.1.	Office refusal
	4.2.	Sampling unit information refused
	4.2.1.	Information refused about number of dwellings/households at address
	4.2.2.	Information refused that would allow identification of desired respondent(s) within dwelling/household
	4.3.	Refusal at introduction/before interview
	4.3.1.	Refusal by desired respondent
	4.3.2.	Refusal by proxy
	4.4.	Refusal during the interview
	4.5.	Broken appointment, no re-contact

Figure 3: Example of a three-level survey outcome-coding frame

⁽¹⁾ Numbering of categories corresponds to the outcomes in Figure 1.

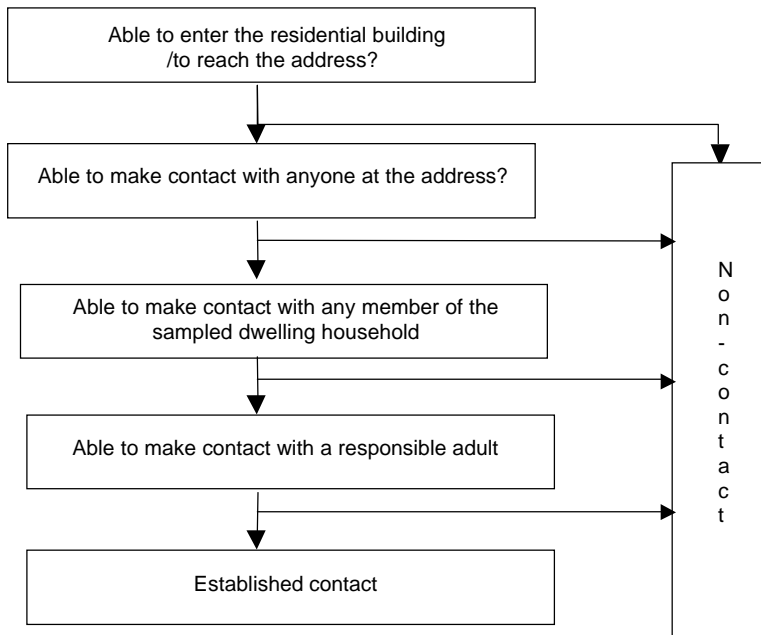


Figure 4: Stages in the contact process

With regard to making contact on household surveys, different situations need to be distinguished. This is especially important when using a list of addresses or dwellings as a sampling frame. Different categories of non-contact should be applied to cases where the eligibility of the address has been established for certain and to cases where it has not (i.e. within first-level categories 3 and 6 in Figure 2). Figure 4 describes the logical stages to contact for a household survey. None of the four stages can be completed until all of the previous stages have been successfully completed, though often stages will be completed simultaneously. Thus, there are four possible types of non-contact, each of which must be recognised in any standard categorisation. It should also be noted that eligibility could be established at any one of the stages in the contact process.

On other types of survey, too, the contact process may have multiple stages. On a business survey, it may be necessary first to make telephone contact (e.g. with a receptionist), then to contact someone who can identify the staff member most likely to be able to supply the requested information, then to contact that person or those persons.

With regard to refusals, the stage reached in the survey process and the status of the person refusing should be recorded. For example, a decision not to participate in the survey is sometimes communicated directly to either the survey organisation or the sponsoring organisation. Such cases are often referred to as ‘office refusals’. Only refusals made before the initial interviewer contact should be categorised as ‘office refusals’. Furthermore, other reasons for non-participation communicated to the office (e.g. illness or language) should be coded to the appropriate subcategory of ‘other non-interview’.

4.2. Outcome rates

4.2.1. Multiple outcome rates

Surveys typically use many outcome categories: more than 15 is common, and up to 30 or 40 is not unheard of. The distribution of sample cases over outcome categories is often summarised in the form of one or more ‘outcome rates’ (ratios). We use the term ‘outcome rates’ as a generic term meaning any ratio calculated from the distribution of cases over outcome categories on a survey.

We believe that there are a number of different outcome rates that can provide useful information for different purposes and that at least some of these rates should be defined in standard ways and reported routinely. These include the four major rates described in AAPOR [1] plus one other and can be loosely defined as follows:

- response rate: the proportion of eligible units for which an interview/questionnaire is completed;
- contact rate: the proportion of eligible units successfully contacted;
- cooperation rate: the proportion of contacted units for which an interview/questionnaire is completed;
- refusal rate: the proportion of eligible units that refuse an interview;
- eligibility rate: the proportion of sample cases found (or estimated) to be eligible.

These general descriptions of the five main types of outcome rate only serve to illustrate the main differences between the rates. In fact, there are a number of alternative definitions for each of the different outcome rates. There may be valid reasons for these alternatives, but in many cases different definitions are used without explicit consideration of the issues. This illustrates the importance of developing detailed definitions and guidelines. Some proposed definitions can be found in Appendix B. Note that we deal here with rates to describe the outcomes with respect to the selected sample. We do not deal here with possible frame under-coverage, though it should be recognised that under-coverage can affect response rates (e.g. if the sub-population excluded from the frame is likely to have a particularly high, or low, survey response rate).

4.2.2. Uses and interpretations

It is important to understand the uses and interpretation of the different rates. For example, eligibility rates can help in planning and design of future surveys that intend to use the same sampling frame. Response rates provide a general indicator of the success of the data collection exercise and of sample coverage. Non-contact rates, refusal rates and cooperation rates — in combination with metadata regarding survey definitions, procedures and constraints — provide information that can help in the development of strategies to improve or maintain response rates in future and also provide indicators of the success of specific aspects of the fieldwork. Furthermore, rates for the total sample can aid

overall assessment and interpretation, while rates for subdomains can be particularly useful for targeting future response-maximisation strategies and for performance monitoring of field offices and interviewers. We propose that standard definitions of outcome rates should be accompanied by guidance on the meaning and interpretation of each different rate. This is conspicuously absent from previous attempts at the definition and/or standardisation of response rates (e.g. [1], [11], [15], [19]).

4.2.3. Component rates

Non-response is a complex phenomenon and knowledge of the response rate alone can hide interesting differences or patterns in the other outcome rates and can lead to incorrect conclusions and inappropriate action being taken. For example, some surveys may have higher refusal rates than others, and the non-contact rate can be exceptionally low due to substantial fieldwork efforts. At the same time, another survey having a similar response rate may have invested more effort on refusal conversion but been less able to minimise the non-contact rate due to a very short fieldwork period. Thus, knowledge of the relevant component outcome rates is needed in order to draw appropriate conclusions about the possible impact of non-response and about future strategies to improve response rates. In Section 5.2 below, we provide an example of this.

4.2.4. Weighted and unweighted rates

Furthermore, the response rate serves at least two important purposes, which have quite distinct implications for the definition of the response rate. The first purpose is as a survey (output) quality indicator. The second is as a fieldwork (process) quality indicator. For the first purpose, the response rate should correctly reflect the structure of the survey population to which estimates pertain. This therefore requires weighting by inverse selection probabilities (design weights) to be used in the response rate calculation. For the second purpose, unweighted response rates can be more appropriate.

We recommend that weighted (by selection probabilities) response rates should always be published. For example, if response rates differ by strata or other intermediate sampling units, which have different inclusion probabilities, unweighted outcome rates could be misleading, both as a measure of process quality and as a measure of survey coverage quality. Also, survey design changes can artificially affect unweighted response weights; only a comparison of weighted rates can give a true reflection of the effect of other factors (survey climate, success of interviewers, etc.) and of any time trends in the contribution of non-response to survey coverage. We give an example of this in Section 5.4 below. Comparison of the weighted and unweighted rates indicates the effect of the survey design. Obviously, any weighting should be described fully, when reporting the response rates [1].

In summary, we would recommend that at least five outcome rates should be calculated and published for each survey. These are the overall response rate, the contact rate, cooperation rate, the refusal rate and the eligibility rate. Additionally, if the sample is not EPSEM, weighted as well as unweighted rates should be published. Moreover, we would suggest that the complete distribution of sample cases over outcome categories — both weighted and unweighted — should be published where possible, to allow sophisticated

users to calculate their own rates. To calculate and publish all these outcome rates may seem excessive, but we believe that each provides different relevant information. Far from adding to confusion about response rates, we believe that the regular appearance of a whole set of outcome rates should alert people to the meanings of each and should help to prevent the inappropriate comparison of rates that have completely different meanings. Indeed, even these 10 rates might be insufficient for some purposes. For example, some surveys currently publish both ‘full’ response rates (where only ‘complete’ responses are counted) and ‘overall’ response rates (where useable partial responses are included too). This too may be an important and useful distinction if some analyses require full response while others can include partial responses too (see Appendix B). However, for the distinction to be useful, a clear definition of full and partial response is needed, as proposed below in Section 4.3.4.

4.3 Survey specific definitions

Each survey should develop and implement explicit definitions of sampling units, rules and procedures for establishing eligibility of sampling units, rules and procedures for collecting data from proxy respondents and definitions of complete response and of acceptable partial response. Each of these definitions should be published in technical reports/appendices alongside response analysis.

4.3.1. Sampling units

Outcomes and response rates must be presented with respect to sampling units^(?). It is therefore important that sampling units are clearly defined. For example, ‘household’ surveys may in fact sample addresses, dwellings, households and/or families, each of which has multiple possible definitions. The choice of definition can affect response rates. For example, a survey of sub-household units may be expected to suffer greater levels of non-contact than a survey of households. Thus, it is important that the definition should be known and taken on board in the interpretation of response rates. We propose that surveys should clearly define, at the design stage, the sampling units with respect to which outcomes will be recorded and presented. The definition should be published in the survey technical report.

4.3.2. Eligibility of sampling units

The definition of eligibility affects the interpretation of response rates. The definition of unknown eligibility affects the calculation of response rates as the calculation should take explicit account of the uncertainty that often surrounds the eligibility of sample units. For example, it is sometimes difficult to be certain whether an address at which no contact has been made is occupied or vacant. Similarly, on a telephone survey of businesses, if a number rings repeatedly with no reply, it may be difficult to know whether the business still exists. Typically, interviewers are forced to make an assumption. This leaves researchers and others no means of taking the uncertainty into account when assessing

^(?) In some cases, it may be appropriate for multi-stage sample surveys to present outcomes separately for each main stage — for example response by schools and response by pupils within schools.

survey outcomes or estimating response rates. Better practice is to allow uncertainty regarding eligibility to be recorded and captured on the data set and subsequently to **estimate** the eligibility rate amongst the cases of uncertain eligibility.

4.3.3. Proxy respondents

Surveys vary in the extent to which, and circumstances in which, proxy responses are allowed. By proxy response we mean the provision of data on behalf of the sample member by someone else (typically, another member of the same household or business). This variation affects response rates. We propose that the definition of the desired respondent(s), the definition of any other acceptable (proxy) respondents and definition of the circumstances in which proxy responses are accepted should all be agreed at the survey design stage, implemented uniformly in the field and reported in the survey technical report. This information can aid interpretation of observed outcome rates.

4.3.4. Complete and partial response

Many surveys make an explicit distinction between ‘complete’ and ‘partial’ response. In practice, all surveys accept some degree of partial information. A partial response is where the sample unit participates in the survey but does not provide all of the desired information. The distinction between a complete and partial interview should be defined and stated explicitly for each survey, as should the distinction between acceptable partial information and unacceptable partial information (unit non-response), as these may vary across surveys. AAPOR [1] suggests some possible definitions. In general, the definitions will depend on the proportion of either applicable or ‘key’ questions that are either answered or administered, or a combination of these. For example, if a key question on employment is not answered and a subsequent block of questions is not asked because of the routing on this key question, the case would be defined as ‘unit non-response’ (specifically, ‘refusal during interview’) rather than a ‘partial interview’.

4.4. Metadata

Other information on aspects of survey design and implementation likely to affect response rates should be published in order to provide contextual information for the interpretation of response rates. Response rates are affected not only by the performance of interviewers, but also by factors like survey-specific definitions, data collection mode, length of the fieldwork period, rules on number of calls made etc. All this should be clearly stated alongside published response rates. In addition, we would propose routine publication of the distribution of number of call attempts for non-contacted units and of number of call attempts until contact was achieved and number of further calls until final outcome was achieved for contacted units.

5. Examples

5.1. Guidance on the distinction between outcome categories

In the course of survey fieldwork, situations of potential ambiguity regarding the appropriate outcome category can arise. One such situation occurs on household surveys when the interviewer is unable to speak to any resident of the sampled dwelling. This situation can be used as an illustration of the sort of guidance that must be provided on the distinction between categories.

If an interviewer does not speak to any resident of the sampled dwelling, then at least three broad outcome categories (at the level of categorisation of Figure 2) are possible:

- non-contact (eligible sample unit),
- vacant (ineligible sample unit),
- eligibility of sample unit uncertain.

(These categories would in practice be divided into subcategories as illustrated in Section 4.1 above, but that is not relevant to the discussion here.)

To be able to choose between these three categories, interviewers must be provided with clear guidance on what constitutes ‘certainty’ regarding whether or not the dwelling is vacant. One could argue that an element of uncertainty will always remain so long as the interviewer has not spoken with any resident, but this is extreme and is almost certainly inconsistent with previous practice on most if not all surveys. More reasonably, interviewers might be provided with guidance on the circumstances in which they are allowed to conclude that a dwelling is vacant, based upon observation alone and the circumstances in which they can accept proxy information, for example from neighbours. We would suggest, for example, that it is reasonable to conclude that a dwelling is vacant if the visible windows and doors (if any) are boarded up or if the interviewer can see through at least two windows and observe a complete absence of furniture and fittings. Clearly, it is still possible that such dwellings might be inhabited, but we would suggest that the likely extent of misclassification would usually be very small. Similarly, we might suggest that a proxy report that the dwelling is vacant should be acceptable provided that the reporter lives in the immediate vicinity, appears knowledgeable and is not intoxicated. Proxy reports can be particularly useful for dwellings such as second homes and holiday accommodation, as these may look occupied but neighbours may know that they are not.

Similarly, guidance will also be required on what constitutes ‘certainty’ that the dwelling is occupied. We would suggest, for example, that observation of any person in the dwelling or observation of vehicles or garden equipment on the property that appear to have been used recently, in combination with the presence of fittings such as curtains, would warrant an assumption that the dwelling is occupied. In practice, guidance might need to be quite detailed to account for different possible combinations of information and observations. It

must also be recognised that ‘certainty’ of occupation does not necessarily imply that the sampling unit is eligible for the survey. For example, the person(s) living there may be there temporarily or part-time, or may fail to meet some other eligibility criterion for the survey. This is why the breakdown into subcategories of non-contact (see Section 4.1) is important in order to be able to estimate outcome rates.

Multi-occupied addresses may pose particular problems for address-based samples, as uncertainty may relate not only to whether or not there is an eligible unit (e.g. household) at the address, but also to the number of eligible units. Guidance would have to deal with these complexities.

5.2. Separation of contact from cooperation

Most surveys quote a single, ‘headline’, response rate. This may mask a number of important differences in the components of response and non-response. For example, two surveys with the same response rate in terms of interviews achieved at eligible households may have different ratios of refusal to non-contacts. The calculation of separate contact and cooperation rates will reveal these differences, as an example from the UK labour force survey (LFS) shows.

Table 1 displays the different outcome rates for three geographic regions in the 1997/98 UK LFS (first quarter). The response rate is defined as the number of completed interviews divided by the number of eligible cases. The contact rate is defined as the number of contacted addresses divided by the number of eligible addresses. The cooperation rate is the number of completed interviews divided by the number of contacted addresses. These are the three regions that had the lowest response rate for that quarter. The headline response rate, which is most commonly reported, tells us something about the relative difficulty of obtaining an interview in these regions. However, it does not reveal whether making contact or obtaining interviews is the main problem. The contact rates show that it is more difficult to make contact in Inner London compared to the other two areas. The cooperation rates show that it is easier to obtain cooperation in Tyne and Wear but that there is little difference between the two London regions once contact has been made. In other words, the difference between Tyne and Wear and the two London regions is mainly due to cooperation rates, while the further difference between the two London regions is mainly due to contact rates.

Table 1: UK labour force survey response rates, contact rates and cooperation rates, first wave 1997/98

	<i>Inner London</i>	<i>Outer London</i>	<i>Tyne & Wear</i>
Response rate	71.0	77.7	80.0
Contact rate	87.4	93.9	92.9
Co-operation rate	81.2	82.7	86.1

By looking at the different components of the response rate in this way we get a better insight into the mechanisms underlying the response process which may help survey organisations to address the distinct causes of non-response.

5.3. Accounting for uncertain eligibility when calculating outcome rates

In many cases, survey outcome codes do not allow for the acknowledgement of any uncertainty regarding whether or not a sample unit is eligible. For example, in face-to-face official surveys in the UK, interviewers have had, until recently, to classify each sample address as either ‘eligible’ or ‘ineligible.’ In reality, eligibility is not always known with certainty, for example in the case of addresses where no contact is made. As non-contacts can account for up to 5 % of sample units on personal interview surveys, this can make some difference to the estimation of response rates (see Appendix B for our proposed definition of response rate). With postal or telephone surveys, the prevalence of cases of uncertain eligibility can be significantly higher.

The RDD component of the 1999 Welsh Assembly election study (WAES) provides an illustrative example. The sample was a simple random sample of telephone numbers [20]. After the exclusion of sample numbers ascertained to be ineligible, uncertainty remained regarding the eligibility of 17 % of numbers. We present here five different estimates of response rate, based upon different estimates of the proportion eligible amongst cases of uncertain eligibility (Table 2). As the proportion estimated to be eligible increases, the estimated response rate decreases. The magnitude of the effect on the response rate is of course relative to the proportion of cases of uncertain eligibility in the sample. In this case, response rate could be anything between 30.3 % and 36.5 %, depending on the assumptions made about the eligibility of cases of unknown eligibility. In fact, a detailed follow-up of these cases was undertaken on the WAES [14] and the eligibility rate amongst this group was estimated to be 19 %, producing an overall estimated response rate of 35.1 %. Another method sometimes used is to apply the eligibility rate observed amongst cases where eligibility was established with certainty. In the case of WAES, this would have led to an estimate of response rate of 32.1 %, perhaps serving as a reminder of the rather arbitrary nature of this approach.

In general, an appropriate method of estimating the eligibility rate amongst cases of uncertain eligibility should be proposed and described in any standard guidance. In some situations it may be reasonable to assume that the proportion eligible amongst those where eligibility is uncertain is the same as the proportion eligible amongst those where eligibility was established [6]. This assumes that the probability of establishing eligibility is independent of eligibility status. In other situations, it may be appropriate to use external (e.g. register) information to estimate the proportion. Ezzati-Rice [5] provides a discussion of methods of estimating eligibility rates and their impacts. More recently, Brick et al. [2] introduced a survival function method which takes account of the number of call attempts made to each sample member. In any case, the sensitivity of the response rate to any estimate of eligibility rate can be demonstrated using the maximum and minimum plausible values.

Table 2: Taking into account unknown eligibility: effect on response rates, WAES 1999

Unknown eligibility ⁽¹⁾		
Assumed eligible	Assumed not eligible	Estimated response rate ⁽²⁾
100 %	0 %	30.3 %
50 %	50 %	33.1 %
19 %	81 %	35.1 %
10 %	90 %	35.7 %
0 %	100 %	36.5 %

⁽¹⁾ Excluding cases where ineligibility was firmly established, eligibility was uncertain for 17.0 % of sample telephone numbers.

⁽²⁾ Source: Welsh Assembly election study 1999.

5.4. Weighted versus unweighted response rates

The importance of weighted outcome rates stems from the possibility that outcome rates could differ across strata or other intermediate sampling units, which have different inclusion probabilities. Data from the British crime survey (BCS) 1996 used to illustrate the effect of (not) weighting. Unweighted response rates are presented in Table 3 for each of the two major sampling strata. The table shows that BCS samples a larger proportion of addresses in inner city areas than other areas. Therefore, we should weight down the response rates for inner cities to correctly reflect the proportion of the population that lives in inner cities.

Table 3: Response to British crime survey 1996

	m_i (responding units)	n_i (units in sample)	N_i (units in population)	n_i/n (proportion of sample units)	N_i/N (proportion of population units)	m_i/n_i (unweighted response rate)
Inner cities	3 868	4 970	2 811 794	25.1	12.9	77.8 %
Other areas	12 480	14 838	18 985 377	74.9	87.1	84.1 %

The unweighted response rate, as reported in the survey technical report (Hales and Stratford 1996), is calculated as follows:

$$RR_U = \frac{\sum_i m_i}{\sum_i n_i} = \frac{16\,348}{19\,808} = 82.5 \%$$

(This can alternatively be formulated as a weighted average of stratum response rates, where the weights are the sample proportions in the strata:

$$RR_U = \sum \left(\frac{n_i}{n} \times \frac{m_i}{n_i} \right) = (.251 \times 77.8) + (.749 \times 84.1) = 82.5 \%$$

Instead, the recommended method is to weight each sample unit by the design weight, namely:

$$RR_w = \frac{\sum_i \frac{N_i}{n_i} m_i}{\sum_i \frac{N_i}{n_i} n_i} = \frac{1}{N} \sum_i \frac{N_i}{n_i} m_i = 83.3 \%$$

(This is exactly equivalent to the alternative formulation as a weighted average of stratum response rates, where the weights are the population proportions:

$$RR_w = \sum_i \frac{N_i}{N} \frac{m_i}{n_i} = (.129 \times 77.8) + (.871 \times 84.1) = 83.3 \%$$

In this particular case, the weighted response rate is slightly higher than the unweighted response rate as the response rate is lower in the over-sampled stratum (inner city areas). The design of the BCS was changed in 2000, so that the survey now over-samples small police force areas. These tend to be relatively rural areas, where response rates are higher than average. Consequently, comparisons of unweighted response rates before and after the design change would not provide valid information about relative field performance, as the overall unweighted response rate would increase even if response rates remained unaltered in every area.

6. Conclusions

Response rates and other outcome rates are potentially very important quality and performance indicators. However, it is clear that much needs to be done to standardise definitions and practice before published rates can be relied upon to convey useful and actionable information and to form the basis of meaningful comparisons. In summary, to make progress towards the standardisation of outcome categories and response rates, we believe that it is necessary to:

- develop a hierarchical schema of standard outcome categories with general applicability to a particular class of survey;
- develop a detailed definition of each category in the schema, with guidance on allocation of sample cases to categories in potentially ambiguous situations;
- define a set of appropriate outcome rates and guidance on calculation of those rates based upon the standard outcome categories;
- provide guidance on the presentation and interpretation of each of these rates;
- define a standard set of metadata that should be presented in association with analysis of survey outcomes and outcome rates;

- provide guidance on how the metadata should influence the interpretation of outcome rates;
- get the standard outcome categories and rates adopted and endorsed by major professional bodies, survey organisations and commissioning and coordinating bodies.

If all of the above were achieved, there would then remain a need continually to monitor and evaluate the implementation of the accepted standards and regularly to re-assess the standards themselves. It should also be noted that appropriate standard definitions, procedures and guidance are likely to vary over broad classes of survey. This must be recognised in any developments of the sort proposed. It may be efficient first to develop definitions, procedures and guidance for just one or two classes of survey and then later extend them to other classes of survey. Classes could be defined by combinations of target population, data collection mode and sampling method, for example in-person interview surveys of individuals, in-person interview surveys of households, RDD telephone surveys, register-based telephone surveys of households or persons, telephone surveys of businesses or other establishments, in-person interview surveys of businesses or other establishments, mail surveys, panel surveys, etc.

In this article, we have made some suggestions as to how these requirements might be met (Section 4). In the UK, we have ourselves initiated a move towards standardisation [12], [13]. In the course of this initiative, we became aware of the extent to which both current practice and our own thinking is influenced by the specific constraints of the sampling frames available and usual survey practices in the UK. Similarly, other initiatives, such as AAPOR [1] have, we believe, naturally been heavily influenced by the circumstances in one particular country. However, the importance of cross-national and international surveys is continuing to grow and the importance of cross-national comparisons of indicators such as response rates is therefore also increasing. We have little doubt, consequently, that the development of generic standards regarding the definition and implementation of survey outcomes and the definition and presentation of outcome rates — with cross-national applicability — is an important challenge whose time has come.

7. Acknowledgements

We are grateful to numerous colleagues at The National Centre for Social Research, London, UK, and the Office for National Statistics, London, UK, for insights into many of the issues discussed in this paper; to members of the ‘International workshop on household survey non-response’ for inspiration and for discussion of our ideas at the Budapest 2000 workshop; to Tom Smith for comments on our draft proposals for standardisation. Opinions expressed in this paper are those of the authors and should not necessarily be attributed to our employing organisations, past or present.

8. References

- [1] AAPOR, *Standard definitions: final dispositions of case codes and outcome rates for surveys*, AAPOR, Ann Arbor, Michigan, 2000.
- [2] Brick, J. M., Montaquila, J. and Scheuren, F. (2002), 'Estimating residency rates for undetermined telephone numbers', *Public Opinion Quarterly*, No 66, pp. 18–39.
- [3] De Heer, W., 'International response trends: results of an international survey', *Journal of Official Statistics*, Vol. 15, 1999, pp. 129–142.
- [4] European Science Foundation, *The European Social Survey (ESS) — A research instrument for the social sciences in Europe*, European Science Foundation, Strasbourg, 1999.
- [5] Ezzati-Rice, T., 'An alternative measure of response rate in random digit dialling surveys that screen for eligible sub-populations', paper presented to the 'International workshop on household survey non-response', Budapest, October 2000.
- [6] Frankel, L., 'The report of the CASRO task force on response rates', Wiseman F (ed.), *Improving data quality in a sample survey*, Marketing Science Institute, Cambridge MA, 1983.
- [7] Groves, R. M., *Survey errors and survey costs*, Wiley Interscience, New York, 1989.
- [8] Groves, R. M. and Couper, M. P., *Non-response in household interview surveys*, John Wiley & Sons, New York, 1998.
- [9] Groves, R. M., Dillman, D. A., Little, R. and Eltinge, J., *Survey non-response*, John Wiley & Sons, New York, 2002.
- [10] Kviz, F. J., 'Toward a standard definition of response rate', *Public Opinion Quarterly*, Vol. 41, 1977, pp. 265–267.
- [11] Lessler, J. T. and Kalsbeek, W. D. (1992), *Nonsampling error in surveys*, John Wiley & Sons, New York.
- [12] Lynn, P., Laiho, J., Martin, J. and Beerten, R., 'A project to standardise response rate estimation in the UK', paper presented to the 'International workshop on household survey non-response', Budapest, October 2000.
- [13] Lynn, P., Beerten, R., Laiho, J. and Martin, J., 'Recommended standard final outcome categories and standard definitions of response rate for social surveys', *Working Papers of the Institute for Social and Economic Research*, Paper 2001-23, University of Essex, Colchester, 2001.

- [14] Nicolaas, G. and Lynn, P., 'Random digit dialling in the UK: viability revisited', *Journal of the Royal Statistical Society — Series A (Statistics in Society)*, No 165, 2002, pp. 297–316.
- [15] Platek, R. and Gray, G. B., 'On the definitions of response rates', *Survey Methodology*, No 12, 1986, pp. 17–27.
- [16] Rydenstam, K. and Wadeskog, A., 'Evaluation of the European time use pilot survey', Eurostat Time Use Surveys Task Force, Document E2/TUS/5/98, Eurostat, Luxembourg, 1998.
- [17] Smith, T., *Standards for final disposition codes and outcome rates for surveys*, NORC/University of Chicago, 2000, <http://www.fcsm.gov/papers/smith.html>.
- [18] Smith, T., 'Developing non-response standards', Groves, R., Dillman, D., Eltinge, J. and Little, R. (eds), *Survey non-response*, Wiley, New York, 2002.
- [19] Statistics Canada, *Standards and guidelines for reporting of non-response rates: definitions, framework and detailed guidelines*, Statistics Canada, Ottawa, 1993.
- [20] Thomson, K., Nicolaas, G., Bromley, C. and Park, A., *Welsh Assembly election study, 1999: technical report*, National Centre for Social Research, London, 2001.
- [21] Hales, J. and Stratford, N., *1996 British Crime Survey (England and Wales): technical report*, National Centre for Social Research, London 1996.

Appendix A — Proposed outcome categories for face-to-face surveys of households ⁽³⁾

Eligible, interview		
1. Complete interview	1.1.	Complete interview by desired respondent(s)
	1.2.	Complete interview: partly by desired respondent and partly by proxy
	1.3.	Complete interview by proxy
2. Partial interview	2.1.	Partial interview by desired respondent
	2.1.1.	<i>Partial household interview</i>
	2.1.2.	<i>Household interview but non-contact with one or more elements</i>
	2.1.3.	<i>Household interview but either refusal or incomplete interview by one or more elements (all elements contacted)</i>
	2.1.4.	<i>Other partial interview by desired respondent(s)</i>
	2.2.	Partial interview: partly by desired respondent and partly by proxy
	2.3.	Partial interview by proxy
	2.3.1.	<i>Partial household interview by proxy</i>
	2.3.2.	<i>Household interview by proxy but non-contact with one or more elements</i>
	2.3.3.	<i>Household interview by proxy but refusal or incomplete interview by one or more elements</i>
	2.3.4.	<i>Other partial interview by proxy</i>
Eligible, non-interview		
3. Non-contact	3.1.	No contact with anyone at the address
	3.2.	Contact made at the address, but not with any member of the sampled dwelling/household
	3.3.	Contact made at the sampled dwelling/household, but not with any responsible resident
4. Refusal	4.1.	Office refusal
	4.2.	Sampling unit information refused
	4.2.1.	<i>Information refused about number of dwellings/households at address</i>
	4.2.2.	<i>Information refused that would allow identification of desired respondent(s) within dwelling/household</i>
	4.3.	Refusal at introduction/before interview
	4.3.1.	<i>Refusal by desired respondent</i>
	4.3.2.	<i>Refusal by proxy</i>
	4.4.	Refusal during the interview
5. Other non-interview	4.5.	Broken appointment, no re-contact
	5.1.	Ill at home during survey period
	5.2.	Away/in hospital throughout field period
	5.3.	Physically or mentally unable/incompetent
	5.4.	Language

⁽³⁾ 'Household' is an ambiguous term as it can mean surveys taking place in households rather than establishments or surveys of households rather than individuals. Although the former is used internationally, especially in NSIs that carry out both sorts of surveys, here we use the term 'surveys of households' in the latter sense. We have also developed a separate categorisation for surveys of individuals (not shown). This schema is suitable for both address-based (area-based samples and samples from address lists) and register-based samples. In the latter case, some categories do not apply.

	5.5. Lost interview 5.6. Other non-response 5.6.1. <i>Full interview achieved but respondent asked for data to be deleted</i> 5.6.2. <i>Partial interview achieved but respondent asked for data to be deleted</i> 5.6.3. <i>Other non-response (give details)</i>
Unknown eligibility, non-interview	
6. Unknown eligibility	6.1. Not attempted 6.1.1. <i>Not issued to an interviewer</i> 6.1.2. <i>Issued but not attempted</i> 6.2. Inaccessible 6.3. Unable to locate address 6.4. Unknown whether address contains residential housing 6.4.1. <i>Information refused about whether address is residential</i> 6.4.2. <i>Unknown whether address is residential due to non-contact</i> 6.5. Residential address — unknown whether eligible household 6.5.1. <i>Information refused about whether there are eligible residents</i> 6.5.2. <i>Unknown whether there are eligible residents due to non-contact</i> 6.6. No screener completed 6.6.1. <i>Refusal to complete screener</i> 6.6.2. <i>Screener not completed due to non-contact</i> 6.7. Other unknown eligibility
Not eligible	
7. Not eligible	7.1. Not yet built/under construction 7.2. Demolished/derelict 7.3. Vacant/empty 7.4. Non-residential address 7.5. Address occupied but no resident household 7.6. Communal establishment/ institution 7.7. Resident household(s) but not eligible for the survey 7.8. Address out of sample 7.9. Other ineligible

NB: Full definitions of these categories can be found in Lynn et al. [13].

Appendix B — Proposed outcome rate definitions for face-to-face surveys of households

Notation used ⁽⁴⁾

<i>RR</i>	=	response rate
<i>COOP</i>	=	cooperation rate
<i>CON</i>	=	contact rate
<i>REF</i>	=	refusal rate
<i>ELIG</i>	=	eligibility rate
<i>I</i>	=	complete interview (1)
<i>P</i>	=	partial interview (2)
<i>NC</i>	=	non-contact (3)
<i>R</i>	=	refusal (4)
<i>O</i>	=	other non-response (5)
<i>UC</i>	=	unknown eligibility, contacted (6.4.1, 6.5.1, 6.6.1, proportion of 6.7)
<i>UN</i>	=	unknown eligibility, non-contact (6.1, 6.2, 6.3, 6.4.2, 6.5.2, 6.6.2 and remainder of 6.7)
<i>NE</i>	=	not eligible (7)
<i>e_C</i>	=	estimated proportion of contacted cases of unknown eligibility that are eligible
<i>e_N</i>	=	estimated proportion of non-contacted cases of unknown eligibility that are eligible

Response rate

We propose that the standard definition of overall response rate should be one that includes partial interviews as respondents. This is why the definition of acceptable partial interview is extremely important (see outcome category 2 and the associated definitions in Lynn et al. [13]). The inclusion of partial interviews should not be used as a means to increase response rates. Rather, partial interviews should be solely those cases that can be used in estimation of at least the key survey estimates.

Another important feature of the proposed response rate definition is that the denominator includes an estimate of the number of eligible non-responding cases amongst those cases where eligibility is uncertain.

$$RR_o = \frac{I + P}{(I + P) + (R + NC + O) + e_C UC + e_N UN}$$

In estimating e_C and e_N one must be guided by the best available objective information and one must not select a proportion in order to boost the response rate. The basis for the

⁽⁴⁾ We have deliberately used the notation of AAPOR [1] as far as possible, to facilitate comparisons of our proposed rates with theirs.

estimate must be explicitly stated and detailed. For some surveys it will be appropriate to assume that the proportion of eligibles amongst those cases where eligibility is uncertain is the same as that amongst cases where eligibility has been established. For other surveys, it will be appropriate to assume $e_C = e_N = 1$. For yet others, different proportions might be assumed for different subcategories of uncertain eligibility.

It will be noted that the response rate is the product of the cooperation and contact rates (defined below).

In addition to the overall response rate, the ‘full response rate’ should also be published. This differs from the overall response rate in that only fully responding cases are counted in the numerator:

$$RR_F = \frac{I}{(I + P) + (R + NC + O) + e_C UC + e_N UN}$$

Cooperation rate

The cooperation rate indicates the number of achieved interviews as a proportion of those ever contacted during the fieldwork period.

$$COOP = \frac{I + P}{(I + P) + R + O + e_C (UC)}$$

Contact rate

The contact rate measures the proportion of all cases in which some household member was reached by the interviewer, even though they might then have refused or been unable to give further information about the household composition or to participate to the survey. Verbal interaction is required to constitute having ‘reached’ someone — leaving a note or an answerphone message is not sufficient.

$$CON = \frac{(I + P) + R + O + e_C (UC)}{(I + P) + (R + NC + O) + e_C UC + e_N UN}$$

In the case of surveys where one person within a household is the target respondent (e.g. a random within-household selection), the proportion of cases where the target respondent was reached by the interviewer may also be of interest.

Refusal rate

In recent years the proportion of refusals has increased significantly on many general population surveys. Therefore it has become increasingly important to monitor refusals separately. If a refusal rate is to be published (though we think this should be optional), we suggest the following definition. The purpose of the refusal rate is to indicate the proportion of all (estimated) eligible cases that refuse. This, in conjunction with the overall response rate, indicates the contribution of refusals to non-response.

$$REF = \frac{R}{(I + P) + (R + NC + O) + e_C UC + e_N UN}$$

Eligibility rate

We propose that the eligibility rate is defined as the ratio of the estimated number of sample cases that are eligible to all sample cases:

$$ELIG = \frac{(I + P) + (R + NC + O) + e_C UC + e_N UN}{(I + P) + (R + NC + O) + (UC + UN) + NE}$$

(An alternative definition often used is the ratio of cases determined to be eligible to cases for which eligibility was determined.)

Effects of interviewers' workload on the Dutch labour force survey

Jan A. van den Brakel

*Department of Statistical Methods, Statistics Netherlands, PO Box 4481,
6401 CZ Heerlen, The Netherlands; e-mail: jbrl@cbs.nl*

Keywords: Embedded field experiments, household survey non-response, interviewers' workload, labour force survey, non-sampling errors

Abstract

A large-scale field experiment embedded in the Dutch labour force survey (LFS) has been conducted to test the effect of interviewers' workload on response rates and the main outcomes of the LFS. During a period of nine months the workload of 70 interviewers, i.e. the number of addresses of the LFS assigned to these interviewers, was varied systematically in a controlled field experiment. It follows that an increase in the interviewers' workload results in a significant increase in the proportion of households that are not visited by an interviewer. The workload level, however, does not influence the response account of the households visited by an interviewer. There are indications that the main outcomes of the LFS are biased, since the visited and not visited households are systematically different with respect to the target parameters of the LFS.

1. Introduction

Response rates to Statistics Netherlands' social surveys declined alarmingly during the 1990s. During the same period, Statistics Netherlands' field staff, who collect data by means of computer-assisted personal interviewing (CAPI), faced increasing capacity problems, resulting in an increase in the number of sample addresses not being visited at all. The question was also raised whether these capacity problems were a significant factor in the increasing non-response rates of the households that were contacted. The increase in non-response rates as well as the number of households that are not visited has a negative effect on the outcomes of the social surveys. First, it results in less precise estimates, since the net sample sizes are reduced (i.e. an increase of the design variance). Second, it might result in less accurate estimates due to selective response (i.e. the estimates are biased). Not only might the non-response be selective, but the households that are not visited might also be selective, since the interviewers decide by themselves, which addresses are visited and which are not. It is not unlikely that interviewers concentrate on neighbourhoods where the highest response rates are expected in case they are not able to visit all the addresses assigned to them.

Interviewers' workload, i.e. the number of sample cases assigned to an interviewer to complete during the data collection period, affects the level of effort of the interviewer applied to contacting and obtaining cooperation from each sampling unit. Therefore, interviewers' workload is one of the design factors of household surveys that influences response rates (Groves and Couper [6], Chap. 10). To shed some light on these issues a

field experiment has been conducted where we systematically varied the workload of approximately 70 interviewers for the Dutch labour force survey (LFS). Possible effects on response rates, the rates of households that were not visited, as well as the main outcomes of the LFS have been investigated. This study naturally leads to an experiment embedded in an ongoing sample survey. Fienberg and Tanur [3], [4], [5], Van den Brakel [12] and Van den Brakel and Renssen [13], [14] discuss the statistical aspects of experiments embedded in sample surveys at length. This paper illustrates the application of this theory in a practical situation.

The objective of this study is described in Section 2. A description of the design of the LFS as well as the experimental design is given in Sections 3 and 4, respectively. Results are presented in Section 5. The major conclusions are summarised in Section 6.

2. Objective

The objective of the experiment is to investigate the influence of the interviewers' workload on response rates and the main outcomes of household sample surveys. We confined ourselves to the field staff collecting data by means of CAPI since the capacity problems mainly occur in this group. All interviewers work in an interview area around their place of residence. Interviewers visit the households living at addresses in their interview area that are drawn in the samples of the different household surveys. Data are collected in face-to-face interviews, supported by electronic questionnaires on a hand-held computer (i.e. CAPI).

In this experiment, workload is defined as the number of addresses per month assigned to an interviewer. Among other factors the total workload of a package of addresses is determined by:

- (i) the time required to complete the questionnaire of a survey;
- (ii) the travel distance between the addresses (this depends on the urbanisation level of the interview area);
- (iii) the average response rates.

The experiment is embedded in the LFS, since this is one of Statistics Netherlands' largest continuously conducted social surveys. This entails that all the interviewers visit addresses to collect data for the LFS each month throughout the year. Therefore, the LFS is an ideal survey to conduct a large field experiment with a group of interviewers that is representative of Statistics Netherlands' field staff.

In the response analysis, we controlled for the workload that an interviewer received from other surveys. For this, the interview time required for the different surveys was taken into account. Simple ratios between the times required to complete a questionnaire of the different surveys and the LFS were derived. The number of addresses for other surveys weighted with these ratios is used as a co-variable in the response analysis.

The objective of this experiment is to test the following hypotheses.

Hypothesis 1: The rate of households that were not visited is not influenced by interviewers' workload, versus the rate of not visited households is influenced by interviewers' workload.

For this purpose, we divided the gross sample into the following groups:

- (i) the number of households visited,
- (ii) the number of households not visited due to high workload,
- (iii) the number of households not visited due to other reasons (illness, vacation).

Hypothesis 2: The response rate of the households that were visited is not influenced by interviewers' workload, versus the response rate of households that were visited is influenced by interviewers' workload.

For this purpose the group of households that were visited is divided into:

- (i) the number of completely responding households,
- (ii) the number of partially responding households,
- (iii) the number of non-responding households (refusal, not at home at least three times, language problems, illness, vacation, special circumstances).

Hypothesis 3: The main outcomes of the LFS are not influenced by interviewers' workload, versus the main outcomes of the LFS are influenced by interviewers' workload.

The following three parameters of the LFS were analysed: the employed labour force, the unemployed labour force and the registered unemployment.

3. Design of the labour force survey

The Dutch labour force survey (LFS) is conducted as a continuous survey, aimed to provide reliable information about the situation on the labour market. The target population of the LFS consists of the non-institutionalised population aged 15 years and over residing in the Netherlands. The sampling frame is derived from a register of all known addresses in the Netherlands. The LFS is based on a stratified two-stage cluster design of addresses. Strata are formed by geographical regions. Municipalities are considered as primary sampling units (PSU) and addresses as secondary sampling units (SSU). In the first stage, a stratified sample of municipalities is drawn with first order inclusion probabilities proportional to the number of addresses. At the second stage, a sample of addresses is drawn without replacement from each selected PSU. Addresses that occur in the register of the employment exchange are over-sampled, because the LFS has to provide accurate outcomes for the monthly publication of the registered unemployment. On the other hand, persons aged 65 and over are under-sampled, since most target parameters of the LFS

concern people aged 15 to 64 years. Principally, all households residing at an address, with a maximum of three are included in the sample. The monthly sample size amounts to approximately 10 000 addresses.

Data are collected by means of CAPI. For all members of the selected households, demographic variables are observed. Target variables are only observed for persons aged 15 years and over. When household members cannot be contacted, proxy-responses are allowed by members of the same household. Households of which none of the members are contacted either directly or by proxy are treated as non-responding households.

The weighting procedure of the LFS is based on the generalised regression estimator. The inclusion probabilities reflect the over- and under-sampling of addresses described above, as well as the different response rates between geographic regions. The weighting scheme is based on a combination of different socio-demographic categorical variables. The integrated method for weighting persons and families of Lemaître and Dufour [9] is applied to obtain equal weights for persons belonging to the same household. Finally, a bounding algorithm is applied to avoid negative weights. Hilbink, Van Berkel and Van den Brakel [7] give a detailed description of the methodology of the LFS.

4. Experimental design

The experiment is designed as a randomised block design (RBD), where interviewers are the block variables. For each interviewer who participated in the experiment, we systematically varied over three different levels of workload, i.e. the number of addresses for the LFS per month. The degree of urbanisation of the interviewer region is a contributory factor to the final workload for a specific number of addresses which are assigned to an interviewer each month. Therefore, the three workload levels in the experiment depended on the degree of urbanisation of the interviewer area. To this end, the interviewers are classified into three different groups based on the degree of urbanisation of their interviewer region. In the first group, more than 50 % of the interview area is highly urbanised. In the second group, more than 50 % of the interview area has a moderate or low degree of urbanisation. The interview areas with a moderate or low urbanisation level are taken together, since the average number of LFS addresses per month in these groups are equal. The third group is a remainder group. The interview areas in this group do not consist of one dominating urbanisation level. These three different groups are indicated as workload area. Within each group, the average number of addresses per month for the LFS (during 1997) was calculated. Each interviewer received three different workloads:

- (i) low workload, i.e. 75% of the average workload,
- (ii) average workload, i.e. the average number of LFS addresses per month of the interviewers' workload area during 1997,
- (iii) high workload, i.e. 125% of the average workload. The three different workloads for the three different workload areas are summarised in Table 1.

Table 1: Number of LFS addresses per month for different workload levels in three different workload areas

Workload area	Treatment		
	Low workload	Average workload	High workload
1	23	31	39
2	17	22	29
3	19	26	33

A random sample of 72 interviewers from Statistics Netherlands' field staff was selected to participate in the experiment. First, the interviewers who met the following requirements were selected:

- (i) at least two years' experience with the LFS,
- (ii) no extremely low or high response rates,
- (iii) no extremely low or high numbers of addresses for the LFS per month.

This resulted in a frame of 382 interviewers. A sample of 72 out of these 382 interviewers was drawn by means of proportional stratified simple random sampling, where the three workload areas are the strata. Each interviewer received each workload level for a period of three consecutive months. As a result, the experiment was conducted in nine months from September 1998 to June 1999. December 1998 was excluded from the experiment. It could be the case that the effect of a change in the monthly workload on response rates becomes manifest only after a certain period of time, since the interviewers must get accustomed to this new workload level. Therefore, each first month of a new workload level in the experiment has been excluded from the analysis.

In order to preclude seasonal effects, interviewers were equally divided over the three different treatment levels at each point in time. This was accomplished by distinguishing three different sequences of workload levels. Within each workload area, interviewers were randomly assigned to one of the three sequences. This design is represented in Table 2. Interviewers, as well as respondents, did not know that they were participating in an experiment to avoid them altering their behaviour, perhaps even unconsciously.

Table 2: Experimental design

Workload area	Number of interviewers	Month								
		1	2	3	4	5	6	7	8	9
1	5	Low workload (23)			Average workload (31)			High workload (39)		
	5	Average workload (31)			High workload (39)			Low workload (23)		
	5	High workload (39)			Low workload (23)			Average workload (31)		
2	15	Low workload (17)			Average workload (23)			High workload (29)		
	15	Average workload (23)			High workload (29)			Low workload (17)		
	14	High workload (29)			Low workload (17)			Average workload (23)		
3	4	Low workload (19)			Average workload (26)			High workload (33)		
	5	Average workload (26)			High workload (33)			Low workload (19)		
	4	High workload (33)			Low workload (19)			Average workload (26)		

The numbers in parentheses are the number of addresses assigned to each interviewer per month for the LFS.

5. Results

The three hypotheses mentioned in Section 2 are tested in this section. The first and second hypotheses concern inferences about the structural relationships between interviewers' workload and the visitation and response account. The conclusions based on these inferences refer to a universal population, which is more general than the finite target population of the Dutch LFS. Therefore, under the assumption that a stochastic model for the visitation and response account can be specified that holds for every observation in the population, a model-based analysis procedure that ignores the sampling design and weighting procedure of the LFS can be applied to test hypotheses about the visitation and response account. The third hypothesis, on the other hand, concerns the treatment effect on estimates of finite population parameters of the LFS. Therefore, a design-based analysis procedure that takes into account the sampling design and weighting procedure of the LFS should be applied in this situation.

During the experiment, seven interviewers dropped out because they stopped working for Statistics Netherlands, or were not available for several months due to illness, vacation or family circumstances. There were no indications that they dropped out due to their participation with the experiment. Finally, 65 interviewers participated for nine months in the experiment. From these 65 interviewers, the addresses of the LFS of the last two months of each workload level were used in the analysis of the experiment. This resulted in a gross sample size of 9 952 households.

5.1. Descriptive results of effects on the visitation and response account

In Table 3 the gross sample is itemised into the number of households visited, the number of households not visited due to higher workload and the number of households not visited due to other reasons (e.g. illness or vacation). It clearly follows that the number of not visited households, increases with the workload.

Table 3: Gross sample itemised to different levels of treatments and visitation accounts

Treatment	Households visited		Households not visited due to workload		Households not visited due to other reasons		Total	
Low	2 215	(91 %)	42	(2 %)	178	(7 %)	2 435	(100 %)
Average	2 905	(87 %)	106	(3 %)	328	(10 %)	3 339	(100 %)
High	3 465	(83 %)	319	(8 %)	394	(9 %)	4 178	(100 %)
Total	8 585	(86 %)	467	(5 %)	900	(9 %)	9 952	(100 %)

Conditional on a household being visited, the effect of the interviewers' workload on their response rates is studied. Therefore, the number of visited households is itemised into the number of fully responding households, partially responding households and non-responding households in Table 4. It seems that, conditional on the visited households, there is no effect of the interviewers' workload on their response rates. The number of

partially responding households is low since proxy-responses are allowed in the LFS (see Section 3).

Table 4: Visited households itemised to different levels of treatments and response accounts

Treatment	Fully responding households		Partially responding households		Non-responding households		Total	
Low	1 093	(49 %)	42	(2 %)	1 080	(49 %)	2 215	(100 %)
Average	1 455	(50 %)	61	(2 %)	1 389	(48 %)	2 905	(100 %)
High	1 699	(49 %)	59	(2 %)	1 707	(49 %)	3 465	(100 %)
Total	4 247	(49 %)	162	(2 %)	4 176	(49 %)	8 585	(100 %)

For each interviewer under each of the three workload levels, a vector is observed that represents the proportion of the visitation account over the three different categories of (i) households visited, (ii) households not visited due to workload and (iii) households not visited due to other reasons. These proportions are based on the total number of households that are assigned to an interviewer under each of the three workload levels during a period of two months. In an equivalent way, for each interviewer under each of the three workload levels a vector is observed that represents the response account of the visited households over the three different categories of (i) fully responding households, (ii) partially responding households and (iii) non-responding households. These proportions are based on the total number of households that are visited by an interviewer during a period of two months under each of the three workload levels.

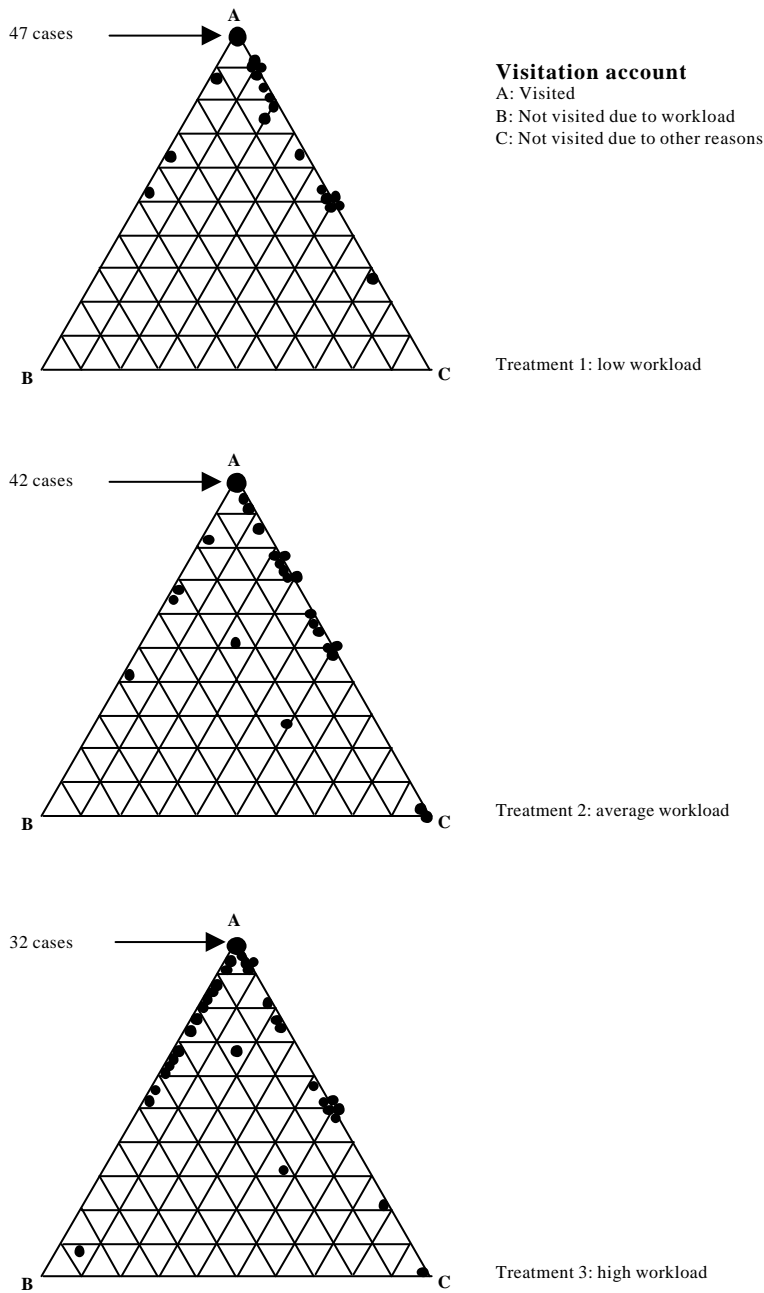


Figure 1: Ternary diagrams of the compositions for visitation account for each workload level.

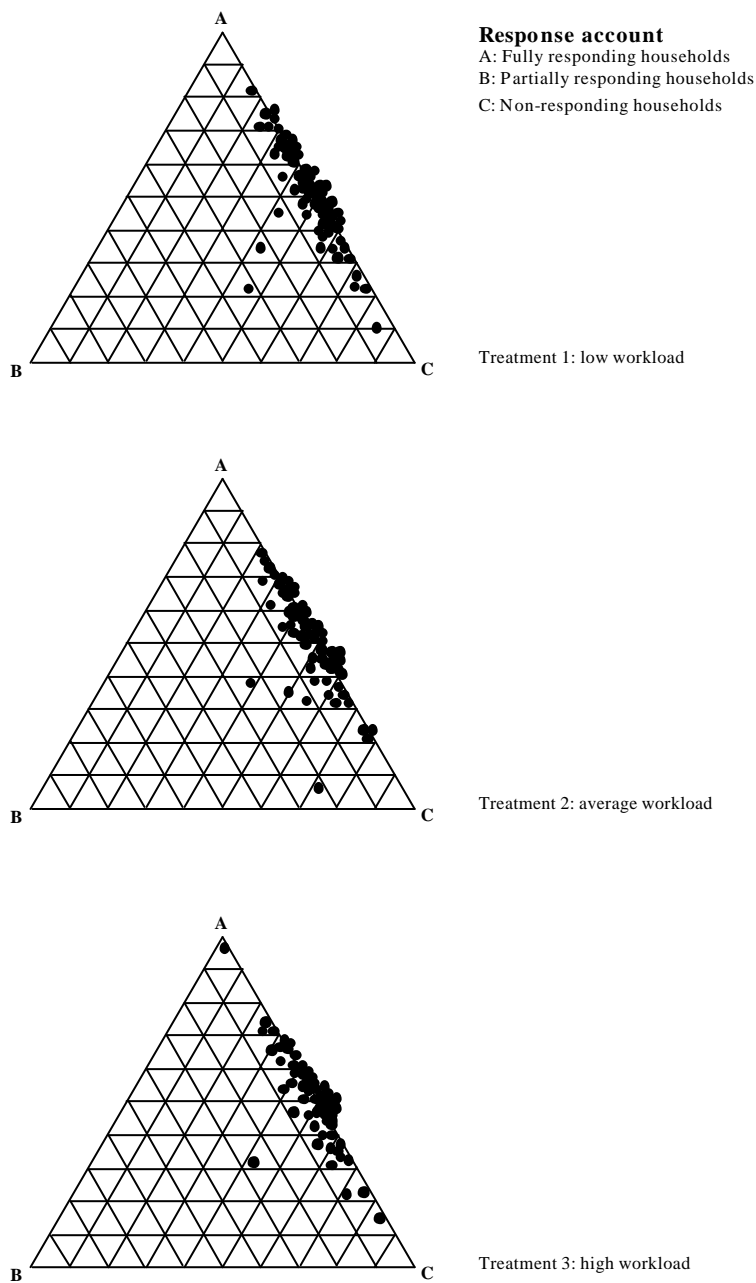


Figure 2: Ternary diagrams of the compositions for response account for each workload level.

These vectors represent the proportional distribution over three categories of a unit and are generally called three-part compositions, Aitchison [2]. The variability of three-part compositions can be displayed by means of a ternary diagram. This is an equilateral triangle with unit altitude. Each vertex corresponds with one of the three parts of the composition and represents the point where the composition completely consists of that part. The side opposite a vertex corresponds with the area where this part does not occur in the composition. The proportion of each part of the composition equals the perpendicular value from a point in the triangle to the side opposite the vertex of that part. The larger the proportion of a component, the nearer the representative point in the triangle lies to its corresponding vertex.

For each workload level, the compositions for the visitation and response account are plotted in a ternary diagram in Figures 1 and 2, respectively. The figures show that the proportion of households not visited increases with the workload level (Figure 1) and that, conditionally on the visited households, the proportions of fully responding, partially responding and non-responding households are not influenced by the workload level (Figure 2).

5.2. Testing the effects on the visitation and response account

5.2.1. Log-ratio analyses

The compositions for the visitation account and the response account can be used to test hypotheses 1 and 2 from Section 2 using a multivariate log-ratio analysis procedure proposed by Aitchison [2]. According to this approach, the log-ratio transformed compositions are analysed using standard multivariate techniques. If the categories of a three-part composition $(x_1, x_2, x_3)^t$ are strictly positive, then the log-ratio transformation is defined as

$$\mathbf{y} = (y_1, y_2)^t = \left(\log\left(\frac{x_1}{x_3}\right), \log\left(\frac{x_2}{x_3}\right) \right)^t,$$

where ‘log’ denotes the natural logarithm. It follows from Figures 1 and 2 for the visitation and the response account that a large part of the compositions contains one or more zero components, which obstructs the application of the log-ratio analysis. It might be possible to separate out the zeros by some form of conditional modelling. The data pattern of the visitation account, for example, can be modelled with a finite mixture for the probability mass on the vertex A and two univariate logistic normal distributions for the data that occur on the A–B and the A–C edges. Then, a likelihood function can be constructed for a compositional data set with such zero components, in order to test parametric hypotheses about treatment effects.

An analysis based on an explicit modelling of the zero components is very laborious. Moreover, it follows directly from the Figures 1 and 2 that workload has a strong influence on the visitation account and nearly no influence on the response account. Therefore a

quicker but less sophisticated approach is applied. Aitchison ([2], Chap. 11) proposed a general procedure which replaces zeros by a positive value, which is smaller than the smallest recordable value. Let δ denote the maximum rounding-off error of the observed compositions. According to this procedure any D-part compositions with C zero and $D - C$ non-zero components is replaced by a composition in which the zeros become $\delta(C + 1)(D - C)/D^2$ and the positive components are each reduced by $\delta C(C + 1)/D^2$. The effect of the zero adjustment on the statistical inference can be evaluated by repeating the analysis for different choices of δ . Aitchison ([2], Chap. 11) proposed to repeat the analysis within a range of $\delta_r/5 \leq \delta \leq 2\delta_r$ with the δ_r the maximum round-off error. In this application $\delta_r = 1\%$ and the analysis is repeated with δ equal to 0.2 %, 0.5 %, 1.5 % and 2 %.

To test hypotheses 1 and 2 (Section 2), the log-ratio transformed compositional data for the visitation and response account are modelled in two multivariate regression models. The following explanatory variables are used:

- (i) intercept;
- (ii) interviewer or block variable (65 levels);
- (iii) workload or treatment (three levels);
- (iv) workload area (three levels);
- (v) time period — a factor with three levels indicating if the composition is observed in the first, second or third set of the three-month period;
- (vi) workload obtained from other surveys — a continuous variable which contains the number of addresses for other surveys, weighted with a ratio that indicates the time required to complete these questionnaires compared with that of a questionnaire of the LFS;
- (vii) experience of the interviewer — a continuous variable which contains the number of months that the interviewer has conducted the LFS.

We started with the following model:

$$\begin{aligned} &\text{intercept} + \text{interviewer} + \text{workload} + \text{workload area} + \text{workload} \times \text{workload area} \\ &+ \text{workload other surveys} + \text{time period} + \text{experience interviewer} \end{aligned} \quad (1)$$

The interaction between workload and workload area as well as the main effect of workload area and experience of the interviewer appeared not to be significant in the models for the visitation and response account. Therefore, the final model to test hypotheses 1 and 2 (Section 2) is given by

$$\text{intercept} + \text{interviewer} + \text{workload} + \text{workload other surveys} + \text{time period} \quad (2)$$

The results of the Manova for the log-ratio transformed compositions of the visitation account and the response account are given in Tables 5 and 6, respectively. The results for the sensitivity analysis are only given for the treatment effects. It follows that the inferences based on these analyses are not sensitive to the choice of δ in the zero

replacement procedure. Workload has a significant effect on the interviewers' visitation account but no significant effect on the interviewers' response account. Workload from other surveys has a significant effect on the visitation and response account. For the response account, this appears to be inconsistent with the non-significant effect of the treatment factor workload. The workload from other surveys and the proportion of visited households, as well as the proportion of fully responding households, are positively correlated. This can be explained since the workload of other surveys appeared to be an interviewer characteristic. Each interviewer has their individually preferred number of sample addresses per month, which is taken into account by dividing the monthly sample addresses over the interviewers. The preferred number of sample addresses per month of better performing interviewers is higher than the poorer performing interviewers.

Since the performance of interviewers on the visitation and response account is strongly determined by their individual characteristics, the interviewer block variable has a strong significant effect on the response and visitation account. Finally, the factor time has a significant influence on the visitation and response account. The percentage of visited households declined during the third period of the experiment, mainly due to vacations of the interviewers. The percentage of fully responding households was slightly lower in the first period of the experiment.

Table 5: Manova for log-ratio transformed compositions of the visitation account

Variable	Wilk's lambda	F (exact)	Hypothesis d.f. ⁽¹⁾	Error d.f. ⁽¹⁾	p-value
Intercept, $\delta = 1\%$	0.040	1 469.953	2	124	0.000
Interviewer, $\delta = 1\%$	0.244	1.981	128	248	0.000
Workload, $\delta = 1\%$	0.858	4.927	4	248	0.001
$\delta = 0.2\%$	0.858	4.930	4	248	0.001
$\delta = 0.5\%$	0.858	4.929	4	248	0.001
$\delta = 1.5\%$	0.858	4.925	4	248	0.001
$\delta = 2\%$	0.858	4.922	4	248	0.001
Workload other surveys, $\delta = 1\%$	0.815	14.093	2	124	0.000
Time period, $\delta = 1\%$	0.730	10.585	4	248	0.000

⁽¹⁾ d.f. stands for 'degrees of freedom'

Table 6: Manova for log-ratio transformed compositions of the response account

Variable	Wilk's lambda	F (exact)	Hypothesis d.f. (1)	Error d.f. (1)	p-value
Intercept, $\delta = 1\%$	0.081	683.081	2	121	0.000
Interviewer, $\delta = 1\%$	0.256	1.844	128	242	0.000
Workload, $\delta = 1\%$	0.990	0.307	4	242	0.873
$\delta = 0.2\%$	0.988	0.354	4	242	0.841
$\delta = 0.5\%$	0.989	0.329	4	242	0.859
$\delta = 1.5\%$	0.990	0.293	4	242	0.882
$\delta = 2\%$	0.991	0.283	4	242	0.889
Workload other surveys, $\delta = 1\%$	0.992	5.122	2	121	0.007
Time period, $\delta = 1\%$	0.906	3.062	4	242	0.017

5.2.2. Alternative analyses

The large number of zeros in the compositions of the visitation and response account hamper the application of the log-ratio analysis. Therefore, several alternative analyses are considered. First, the nonparametric test of Friedman for complete block designs (Lehmann [8], Chap. 6) is applied to test the hypotheses of workload effects on visitation and response account. Three separate tests are applied to test the effect of workload on the three categories of visitation account. Interviewers are the block variables. The proportion of households of the corresponding category under the three different workload levels within each block, i.e. the components of the compositions obtained from the interviewers, are the data to be analysed. Similarly, three separate tests are applied to analyse the effect of workload on the three different categories of response account. The major drawback of this approach is that three univariate tests are applied for the analysis of one multivariate problem. The results of these tests agree with the conclusions obtained with the log-ratio analyses, i.e. workload has a significant effect on the visitation account but no significant effect on the response account.

In addition, multinomial logit models (Agresti [1], Chap. 9) are considered to model the visitation and response account, using households as sampling units. The visitation results for each household specified as a categorical variable over the categories visited, not visited due to workload and not visited due to other reasons are the dependent variables in a multinomial logit model. Similarly, the response results for households specified as a categorical variable over fully responding, partially responding and not responding are the dependent variables in a multinomial logit model. Model 1 is applied to describe these multinomial responses. Since interviewers are incorporated as block variables in these models, many cells with an expected frequency of zero or one arise. This results in severe convergence problems in estimating these models and obstructs the application of a multinomial logit model to analyse the effect of interviewers' workload on the visitation and response account.

Finally, logistic regression models are considered for modeling the visitation and response account. Two logistic regression models for the binary response of households are

estimated to test hypotheses about the visitation account. In the first logistic regression model, the binary response for each household over the two categories visited versus not visited is used as the dependent variable. In the second logistic regression model, the binary response not visited due to workload versus its complement visited or not visited due to other reasons is used as the dependent variable. Similarly, two logistic regression models for the binary response of households are estimated to test hypotheses about response account. In the first logistic regression model, the binary response for each household over the two categories fully responding versus partially or non-responding is used as the dependent variable. In the second logistic regression model, the binary response partially responding versus its complement fully responding or non-responding is used as the dependent variable. Model 1 is applied to describe these binary responses. The convergence problems encountered with the multinomial logit models were solved, since, in each binary response model, two of the three categories are collapsed. This approach, however, has two major disadvantages. First, two univariate tests are applied for one multivariate analysis problem. Second, in the experiment, clusters of households assigned to an interviewer during a period of two months are randomised over the treatments. Therefore, the application of a logistic regression model for the binary response of households will overestimate the significance of the treatment effects. Nevertheless, the results of these tests agree with the conclusions obtained with the log-ratio analyses, i.e. workload has a significant effect in the two logistic regression models of the visitation account but no significant effect in the two logistic regression models for the response account.

5.3. Effects on outcomes of the labour force survey

From the analyses in the preceding section, it follows that a higher workload results in an increase of the proportion of not visited households. If the households not visited are a selective group with respect to their target parameters of the LFS, then the interviewers' workload might influence the main outcomes of the LFS. Therefore, hypothesis 3 from Section 2 is tested. Based on the three subsamples in the experiment, estimates of the parameters of the LFS observed under the three different workloads of the interviewers can be obtained.

The purpose of this analysis is the estimation of the main parameters of the LFS under the different treatments and to test hypotheses about the differences between these estimated population parameters. The application of model-based procedures that do not allow for the sampling design and the weighting procedure of the LFS, might result in design-biased parameter and variance estimates. Consequently, the results obtained with such a model-based procedure might be incommensurable with the parameter and variance estimates of the regular LFS. In order to test hypotheses concerning differences between finite population parameters, Van den Brakel [12] derived a design-based analysis procedure for experiments embedded in sample surveys. In this approach, design-unbiased estimators are derived, e.g. the Horvitz–Thompson estimator or the generalised regression estimator, for the population parameters as well as the covariance matrix of the contrasts between these estimators under the joint probability structure of the sampling design and the experimental design. This naturally leads to a design-based Wald's statistic that draws inference about

the finite population parameters of the sample survey. Technical details of this procedure are given in the appendix.

The employed labour force, the unemployed labour force and the registered unemployment are the main parameters of the LFS analysed in this experiment. These parameters are expressed as percentages of the population aged 15 to 64 years. Results are given in Table 7.

Table 7: Analysis results for three parameters of the Dutch LFS

Parameter	Workload			Wald's statistic	Degrees of freedom	<i>p</i> -value
	low	average	high			
Employed labour force	66.1	60.9	63.4	9.15	2	0.01
Unemployed labour force	3.28	3.67	2.79	3.06	2	0.22
Registered unemployment	2.13	2.23	1.68	3.28	2	0.19

Only the three estimates of the employed labour force are significantly different at a significance level of 0.05. It is not clear why the employed labour force for the average and high workload is lower with respect to the low workload. The observed differences might be a result of selectivity in the households that are not visited by the interviewer. In the preceding section, we saw that the proportion of households that are not visited increases with the workload. Some selectivity in the outcomes of the LFS might be introduced, since interviewers decide by themselves which households are visited and which not. It might also be an artefact, however, since the observed effect of workload on the employed labour force is not monotone. Practical consequences for the organisation of the fieldwork are discussed in the conclusions in Section 6.

6. Conclusions

In this experiment, we first investigated the effect of the interviewers' workload on their visitation account, i.e. the proportion of households visited, not visited due to workload and not visited for other reasons. We also investigated the influence of interviewers' workload on the response account of the households that were visited, i.e. the proportion of fully, partially and non-responding households. For the visitation account, it follows that the proportion of households that were visited declines if the interviewers' workload increases. It appears that the current organisation of Statistics Netherlands' fieldwork allows interviewers to decide how many addresses they visit per month. If the number of addresses assigned to an interviewer exceeds the limit of the interviewers' personal capacities, then this results in an increase in the number of households that are not visited. The response account of the households that were visited, however, seems not to be influenced by the interviewers' workload. There are indications that the target parameters of the LFS of households that are not visited are systematically different from those of households that are visited. As a result, this interviewers' selection mechanism has an

undesirable influence on the outcomes of the LFS and probably also on outcomes of other surveys.

From these results, it follows that an improvement of the response rates cannot be expected from a reduction of the interviewers' workload. The capacity problems of Statistics Netherlands' field staff, however, resulted in an undesirable large proportion of households not visited. Therefore, the planning system for the fieldwork has been improved in such a way that the available capacity of the field staff is known at a detailed regional level. Based on this information, the sample sizes of the surveys can be adjusted if necessary. This should minimise the proportion of not visited households and the accompanying negative effects on the outcomes of our surveys. Reducing the original sample sizes until they match with the available interviewer capacity, however, is an unsatisfactory and temporary solution. In order to settle these capacity problems more permanently, the freelance status of Statistics Netherlands' field staff is changed to permanent contracts. This should make the fieldwork more attractive, which enables Statistics Netherlands to recruit more field staff who are obligated to visit a minimum number of households during the data collection period of a survey.

7. Acknowledgements

The author wishes to thank Marion Hofmans and Mariëtte Vosmer for their valuable contributions and pleasant cooperation during the conduct of this experiment. He also wishes to thank Paul Smith (ONS) and the unknown referees for commenting on an earlier version.

The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

8. References

- [1] Agresti, A., *Categorical data analysis*, John Wiley & Sons, New York, 1990.
- [2] Aitchison, J., *The statistical analysis of compositional data*, Chapman & Hall, London, 1986.
- [3] Fienberg, S. E. and Tanur, J. M., 'Experimental and sampling structures: parallels diverging and meeting', *International Statistical Review*, Vol. 55, No 1, 1987, pp. 75–96.
- [4] Fienberg, S. E. and Tanur, J. M., 'From the inside out and the outside in: combining experimental and sampling structures', *The Canadian Journal of Statistics*, Vol. 16, No 2, 1988, pp. 135–151.
- [5] Fienberg, S. E. and Tanur, J. M., 'Combining cognitive and statistical approaches to survey design', *Science*, Vol. 243, 1989, pp. 1017–1022.

- [6] Groves, R. M. and Couper, M. P., *Non-response in household interview surveys*, John Wiley & Sons, New York, 1998.
- [7] Hilbink, K., Van Berkel, C. and Van den Brakel, J.A., 'Methodology of the Dutch labour force survey, 1987–1999'. research paper, BPA No 2297-00-RSM, Department of Statistical Methods, Statistics Netherlands, Heerlen, 2000.
- [8] Lehmann, E. L., *Nonparametrics: statistical methods based on ranks*, McGraw-Hill, New York, 1975.
- [9] Lemaître, G. and Dufour, J., 'An integrated method for weighting persons and families', *Survey Methodology*, Vol. 13, 1987, pp. 199–207.
- [10] Nieuwenbroek, N. J., 'An integrated method for weighting characteristics of persons and households using the generalised regression estimator', research paper, BPA No 8445-93-M1, Department of Statistical Methods, Statistics Netherlands, Heerlen, 1993.
- [11] Särndal, C. E., Swensson, B. and Wretman, J., *Model assisted survey sampling*, Springer-Verlag, New York, 1992.
- [12] Van den Brakel, J. A., 'Design and analysis of experiments embedded in complex sample surveys', Ph.D thesis, Erasmus University of Rotterdam, 2001.
- [13] Van den Brakel, J. A. and Renssen, R. H., Design and analysis of experiments embedded in sample surveys, *Journal of Official Statistics*, Vol. 14, No 3, 1998, pp. 277–295.
- [14] Van den Brakel, J. A. and Renssen, R. H., 'Analysis of experiments embedded in complex sampling designs', research paper 0110, BPA No 10801-01-TMO, Department of Statistical Methods, Statistics Netherlands, Heerlen, 2001.

Appendix: A design-based analysis procedure for embedded experiments

From the third hypothesis mentioned in Section 2, it follows that the purpose of this experiment is estimating the main parameters of the LFS under the three different workload levels of the interviewers, and testing hypotheses about the differences between these estimates. Therefore, a design-based procedure is applied that takes into account the randomisation mechanism of the experimental design, the sampling design of the LFS and the weighting procedure of the LFS. To this end, a generalised regression estimator for the parameters of the LFS under the different treatments of the experiment and the covariance matrix of the differences between these parameter estimates is derived under the joint probability structure of the experimental design and the sampling design of the LFS. This gives rise to a design-based Wald's statistic to test hypothesis 3 from Section 2.

Let \bar{Y}_1 , \bar{Y}_2 and \bar{Y}_3 denote a parameter of the LFS based on data collected with interviewers assigned to the low, average and high workloads, respectively. Now, the hypothesis of no treatment effects in the parameters of the LFS can be formulated more formally by

$$H_0 : \bar{Y}_1 = \bar{Y}_2 = \bar{Y}_3$$

$$H_1 : \bar{Y}_k \neq \bar{Y}_{k'}, k, k' = 1, 2, 3 \text{ and } k \neq k' \text{ for at least one pair.} \quad (\text{A.1})$$

Let s_k denote the subsample assigned to treatment k . Let B denote the number of blocks in the experiment and m_{bk} the number of responding households in block b assigned to treatment k . Let $m_{+k} = \sum_{b=1}^B m_{bk}$ denote the number of responding households in subsample s_k , $m_{b+} = \sum_{k=1}^3 m_{bk}$ the number of responding households in block b and $m_{++} = \sum_{b=1}^B \sum_{k=1}^3 m_{bk}$ the number of responding households in the entire sample. Let y_{ijk} denote the observation obtained from person i belonging to household j assigned to treatment k . Then $y_{jk} = \sum_{i=1}^{n_j} y_{ijk}$ is the observation obtained from household j assigned to treatment k with n_j the number of persons in household j aged 15 years and over. Furthermore, let \mathbf{x}_j denote an H -vector with the household totals of the auxiliary variables of the persons aged 15 years and over of the j th household.

Each subsample can be considered as a two-phase sample where the design of the survey sample and the experimental design determine the designs of the first and second phases, respectively. Then, the generalised regression estimator for the population mean \bar{Y}_k based on the observations obtained from the households in subsample s_k is defined by

$$\hat{Y}_{reg;k} = \hat{Y}_k + \hat{\mathbf{b}}_k^t (\bar{\mathbf{X}} - \hat{\mathbf{X}}_k). \quad (\text{A.2})$$

Here

$$\hat{Y}_k = \frac{1}{N} \sum_{b=1}^B \sum_{j=1}^{m_{bk}} \frac{m_{b+}}{m_{bk}} \frac{y_{jk}}{p_j}$$

denotes the Horvitz–Thompson estimator of the population mean \bar{Y}_k , and N denotes the size of the target population of the LFS. In (A.2) $\bar{\mathbf{X}}$ denotes a vector of order H containing the population means of the auxiliary variables. Furthermore, $\hat{\mathbf{X}}_k$ denotes the Horvitz–Thompson estimator for $\bar{\mathbf{X}}$ based on the m_{+k} households in subsample s_k , which is given by

$$\hat{\mathbf{X}}_k = \frac{1}{N} \sum_{b=1}^B \sum_{j=1}^{m_{bk}} \frac{m_{b+}}{m_{bk}} \frac{\mathbf{x}_j}{p_j}.$$

An estimator for the regression coefficients of the generalised regression estimator, based on the m_{+k} households in subsample s_k is given by

$$\hat{\mathbf{b}}_k = \left(\sum_{b=1}^B \sum_{j=1}^{m_{bk}} \frac{m_{b+}}{m_{bk}} \frac{\mathbf{x}_j \mathbf{x}_j^t}{w_j^2 p_j} \right)^{-1} \sum_{b=1}^B \sum_{j=1}^{m_{bk}} \frac{m_{b+}}{m_{bk}} \frac{\mathbf{x}_j y_{jk}}{w_j^2 p_j}.$$

Here w_j^2 denotes the model variance of the generalised regression estimator. The first order inclusion probabilities p_j are based on the over-sampling of addresses which occur in the register of the employment exchange, the under-sampling of addresses with only persons aged 65 years and over, and the reduced sample size in July and August.

The weighting procedure is conducted at the level of household totals. The model variance of the generalised regression estimator is chosen proportional to the size of the household, i.e. $w_j^2 = n_j w^2$. Nieuwenbroek [10] shows that in this situation the weighting procedure corresponds to the integrated method for weighting persons and households proposed by Lemaître and Dufour [9]. The following weighting scheme, which contains the most important auxiliary information from the regular weighting scheme of the LFS, was applied in the analysis of this experiment:

age + sex + marital status + region,

where the four variables are categorical.

To test hypothesis (A.1), Van den Brakel [12] derived the following design-based Wald's statistic

$$W = \sum_{k=1}^K \frac{\hat{\bar{Y}}_{greg;k}}{\hat{d}_k} - \left(\sum_{k=1}^K \frac{1}{\hat{d}_k} \right)^{-1} \left(\sum_{k=1}^K \frac{\hat{\bar{Y}}_{greg;k}}{\hat{d}_k} \right)^2$$

where $K=3$, and \hat{d}_k are the variance components defined by

$$\hat{d}_k = \sum_{b=1}^B \frac{1}{m_{bk}} \frac{1}{(m_{bk} - 1)} \sum_{j=1}^{m_{bk}} \left(\frac{m_{b+} (y_{jk} - \hat{\mathbf{b}}_k^t \mathbf{x}_j)}{Np_j} - \frac{1}{m_{bk}} \sum_{j=1}^{m_{bk}} \frac{m_{b+} (y_{jk} - \hat{\mathbf{b}}_k^t \mathbf{x}_j)}{Np_j} \right)^2.$$

It is assumed that a finite population central limit theorem holds, such that the estimates of a population parameter under the three different treatments are multivariate normally distributed. Then the Wald's statistic is, under the null hypothesis of no treatment effects, asymptotically distributed as a chi-squared random variable with $K - 1$ degrees of freedom.

Statistical research at Statistics Norway

Johan Heldal, Jan Bjørnstad, Anne Gro Hustoft, Dinh Q. Pham, Dag Roll-Hansen and Li-Chun Zhang

Division for Statistical Methods and Standards, Statistics Norway

1. Introduction

The very first attempt on research in statistical methodology in Statistics Norway (SN) took place in the 1890s when the Bureau, under leadership of Anders N. Kiær, in 1894 carried out the first social science sample survey ever that aimed at producing estimates generalised to an entire national population. The sample consisted of 80 000 'representative adult men and women'. The background for the survey was the need for statistical information for a parliamentary commission on disablement and age insurance, which prepared new legislation in the field. The 1894 survey was preceded by an empirical study comparing samples of the 1890 census forms with the results of the census. This study gave sufficiently satisfactory results to signal a go-ahead for the real survey. The survey was presented at the ISI meeting in Bern in 1895. It was heavily criticised, but Kiær succeeded in getting his representative method on the ISI agenda for the next four ISI meetings as well. However, the resistance against the method persisted and, without a proper theoretical foundation, Kiær was unable to defend the method in the long run. After 1903, the debate faded away. No more sample surveying was carried out. Another 30 years would pass before the needed foundation was given by Neyman [6].

Research on statistical methods in Statistics Norway was not revived until the 1970s. Starting with a survey of living conditions in 1958, SN had already carried out a number of sample surveys and needed a professional group of statisticians to work out efficient designs and estimation methods for the surveys. In 1975, the Method Unit was founded with a few members and headed by Ib Thomsen. The members of the unit published research on various problems in official statistics, ranging from survey methodology and small area estimation through environmental and economic statistics. Some of the works are listed in the references.

There are other research units in Statistics Norway dealing with economic and econometric analysis, analysis of demographic changes and causes, and population forecasts. Although this research relies heavily on data from SN, this type of research does not typically take place within national statistical offices. For this reason they will not be dealt with in this paper.

Today, the unit responsible for most of the methodological research in Statistics Norway is called the Division for Statistical Methods and Standards and has expanded to about 20 persons working in various fields.

These fields can be grouped as:

- general survey design and estimation,
- non-response and imputation,
- variance estimation,
- small area estimation and registers,
- census methodology,
- time series and seasonal adjustment,
- questionnaire design,
- disclosure control,
- statistical standards and metadata.

The next sections give an overview over recent, ongoing and planned work within some of these fields.

2. Non-response and imputation

In all sample surveys we have the problem of non-response, resulting in incomplete data. As in many other countries, the response rates in Norway have shown a declining trend. Therefore, the Division for Statistical Methods and Standards is doing research on developing model-based methods for reducing the bias due to non-response, mainly weighting adjustment for unit non-response and imputation for item non-response. The present imputation routines are not model-based and will therefore necessarily be unsatisfactory in many cases. The model-based imputation approach is based on stochastic models for the response mechanism, typically logistic regression, and a population model. The basic imputation method considered is random draws from the estimated distribution given no response.

An important research issue is to develop measures of uncertainty that take into account non-response (possibly non-ignorable) and the imputation method used — a general approach for estimating the sample variance in social surveys when the sampling fraction is small has been developed, based on bootstrapping (see [10]).

3. Variance estimation

Statistics Norway plans to increase its efforts in giving estimates of uncertainty in official statistics, taking into consideration sampling design, non-response and imputation method. As mentioned in Section 2, we are currently doing research on estimating the sample variance.

The research is aimed more generally at developing variance estimates and related confidence intervals, of different types, not only the usual estimated sample variance (SV) of the population total estimator. Other measures of uncertainty are estimated conditional SV for poststratified and calibrated estimators, estimated model variance and estimated method variance. The first three measures are well known. To describe the uncertainty aspect indicated by the method variance, consider planning repeated surveys like the labour force sample survey, where it is important to evaluate how the estimator behaves in the long run. For variation, this can be measured by the estimated method variance, based on both the sampling design and a population model. The estimated sample variance is often used for this purpose, mistakenly.

The different variance measures are studied from the perspective of the generalised likelihood principle as formulated in Bjørnstad [1]. This principle is used as a guide for how to make a scientific evaluation of measuring uncertainty in an estimate for **a given sample**, leading to modelling approaches.

4. Small area estimation and registers

4.1. Small area estimation

Small area estimation has been the subject of a number of studies at SN throughout the years. A wide range of methods has been investigated including synthetic estimation, empirical Bayes methods, model-based approach and neural network. We refer to Zhang [9] for a recent report on these experiences.

Current research is connected with the Eurarea project (IST 2000 26290) under the EU's fifth research framework, where Statistics Norway participates together with the Office for National Statistics (UK), Statistics Finland, Istituto Nazionale de Statistica (ISAT, Italy), Statistics Sweden and Instituto Nacional de Estadística (INE, Spain). These are joined with the academic researchers at the University of Southampton (UK), Jyväskylä (Finland) and Poznań (Poland). The project is contracted for the three-year period 2001–03. It has four major themes, namely, use of time series data, use of geo-spatial information, survey data with complex sample design, and small area cross-classifications.

Statistics Norway has undertaken the theoretical development for estimation of small area cross-classifications. The proposed methods will be tested based on the data from Sweden and Italy. Properties of the alternative approaches will be compared based on a large-scale simulation study. Outcomes of the project will include documentation of the related theories, recommendation for practice and software packages for implementation.

4.2. Use of administrative registers

Use of administrative registers often improves surveys by reducing the sampling variance, reducing the bias caused by non-coverage and non-response, and imposing consistency between the various sources of data. Thomsen and Holmøy [7] review some of the previous works at Statistics Norway.

Instead of considering single surveys, recent research has focused on the effects of registers for measure of changes, see Thomsen and Zhang [8]. The current research aims at methods beyond post-stratification and calibration, which may be more suitable and flexible for certain types of data, such as household income and wealth. Research in this respect is closely linked to that on combining data sources and integrated statistics.

5. Time series and seasonal adjustments

We use the X-12 ARIMA seasonal adjustment program developed by the time series staff of the Census Bureau's Statistical Research Division to make seasonal adjustments of economic time series. We have had to make add-ons to make correct adjustments for Norwegian holidays. We have also made a program to simulate the revision for the seasonally adjusted value of the last observation when the value of the following month is observed. The program TRAMO/SEATS will be used to test the series of foreign trade.

In cooperation with the Division for Social and Economic Research and the University of Oslo, ARIMA models for forecasts of the total fertility rate, life expectancy at birth, and migration have been developed (Keilman and Pham, 2000 [5]).

6. Questionnaire testing and design

Statistics Norway has five persons, working mainly with research, development and counselling and giving courses connected to questionnaire measurement design, constituting the Group for Survey Methodology. The research is currently focused on studies of response burden, non-response and response quality, in the development of new survey instruments and in studies of instrument effects. Some examples of the activities in the group are listed below.

Non-response and quality of response in the 2001 census

In 2001 the Norwegian census was held. The Group for Survey Methodology has been closely involved in the work with the census by developing the questionnaire, both a paper version and an electronic version. Since the respondent could choose between a paper version and an electronic version of the questionnaire, the effects of non-response, response burden and quality of response between the two questionnaires will be studied.

Respondent burden in Internet surveys

Different qualitative methods have been applied in order to involve future respondents in the development of Internet questionnaires. At present we focus on what we have learned about the response burden in questionnaires directed towards institutions and businesses. Most of the tests we report from have been conducted as a mixture of individual usability tests and cognitive interviewing carried out at the respondents' place of work.

The 2001 workshop

The Group for Survey Methodology organised the 12th international workshop on household survey non-response, last year. At this conference, Trine Dale and Bengt Oscar Lagerström [2] presented the paper 'The effect of interviewers' attitudes on their work results', where some preliminary results from an interview survey conducted the same year were presented.

Participation in the development of reporting systems

We participate in the development of a system for electronic reporting of information from businesses to the government. Recently we had a central role in revising an Internet-based reporting system for collecting information needed by the government from primary schools.

7. Disclosure control

Disclosure control is relatively new as a research topic in Statistics Norway. Work done so far has been inspired by the disclosure risks in micro-data sets released to researchers, and on easily available register variables that are often linked to these data sets in particular (see Heldal [3], [4]). There has also been some work related to tables. Methods for stochastic controlled rounding have been extended to multiple two-way marginals for higher dimensional tables (unpublished). The 'Web statistics data bank' project in SN, which will be launched by SN this year, will pose new research challenges in this field.

8. Statistical standards and metadata

8.1. General aspects

Work concerning statistical standards in Statistics Norway is decentralised so that each division is given the responsibility for classifications within their own statistical field. This concerns all aspects of the classifications unless otherwise decided by the director-general.

For the purpose of coordination, a Standards Committee has been established. The committee acts as a catalyst and supervises statistical standards. Our Division for

Statistical Methods and Standards serves as the secretariat of the Standards Committee and is responsible for the central work connected to standards.

8.2. Classification database

Statistics Norway, in cooperation with Statistics Denmark, is establishing a database for standard classifications (Stabas). Stabas will also be a tool for the production of statistics. In connection with this work, Statistics Norway has participated in the Neuchâtel group (also consisting of Statistics Sweden, Statistics Denmark, Statistics Switzerland and Run-Software from Germany), where the aim has been to establish a common terminology for classification databases.

8.3. Metadata

Another important part of our standardisation work is connected to metadata. We have started developing a system where all our important/central variables will be documented. This will be used as a tool for standardisation of variables, and will improve accessibility to variable information for all users inside (and eventually outside) Statistics Norway. We regard this system as a first step in an effort to coordinate and link the different metadata systems in Statistics Norway.

Statistics Norway is also developing a dissemination database for aggregate data. The system is called The Statistics Bank and is developed in cooperation with Statistics Sweden and Statistics Denmark. Structured metadata is an important part of the database. In addition, Statistics Norway participates in several metadata projects in the EU's fifth framework programme.

9. References

- [1] Bjørnstad, J. F., On the generalisation of the likelihood function and the likelihood principle, *Journal of the American Statistical Association*, Vol. 91, 1996, pp. 791–806.
- [2] Dale, T. and Lagerström, B. O., 'The effect of interviewers' attitudes on their work results'.
- [3] Heldal, J., 'Confidentiality problems related to survey data in Norway and some possible approaches', Working Paper 41, joint ECE/Eurostat work session in confidentiality, 2001a.
- [4] Heldal, J., 'A ranking approach to confidentiality in survey data', *Proceedings of the annual meeting of the American Statistical Association*, 5–9 August 2001, 2001b.

- [5] Keilman, N. and Pham, D. Q., Predictive intervals for age-specific fertility, *European Journal of Population*, Vol. 16, 2000, pp. 41–66.
- [6] Neyman, J., ‘On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection’, *Journal of the Royal Statistical Society*, Vol. 97, 1934, pp. 558–625.
- [7] Thomsen, I. and Holmøy, A. M. K., ‘Combining data from surveys and administrative record systems: the Norwegian experience’, *International Statistical Review*, Vol. 66, 1998, pp. 201–221.
- [8] Thomsen, I. and Zhang, L.-C., ‘The effects of using administrative registers in economic short-term statistics: the Norwegian labour force survey as a case study’, *J. Off. Statist.*, Vol. 17, 2001, pp. 285–294.
- [9] Zhang, L.-C., ‘Some Norwegian experience with small area estimation’, *Statist. Trans.*, Vol. 4, 2000, pp. 649–664.
- [10] Zhang, L.-C., ‘A method of weighting adjustment for survey data subject to non-ignorable non-response’, Discussion Paper 311, Statistics Norway, 2001.

Note to authors

ROS welcomes contributions from authors on results of research activities in official statistics. Contributions will normally be accepted in English. Nevertheless, reports in any other official languages of the European Union will be considered for publication, subject to the author submitting a summary of not more than 200 words in English. This summary must be submitted to The Executive Editor (at the address below) at the same time as the paper.

Before submitting their papers, authors are advised to seek assistance in the writing of their papers for the correct use of English.

Copyright: In submitting a paper, the author implies that it contains original unpublished work which has not (and is not planned to be submitted) for publication elsewhere. If this is not the case and the paper has been submitted elsewhere for publication, or actually already published, the author must clearly indicate this on the first page.

Pre-assessment: A first evaluation of each paper will be done as soon as possible and authors will be informed of this within a few weeks of the submission. Accepted papers will be published within six months of the author approving the final proof.

Submission format: The author should submit only one copy of his manuscript on paper. This should be accompanied by a summary of not more than 100 words. Manuscripts should in addition be sent electronically - that is, on diskette or by electronic mail. This will facilitate the editing process.

If a diskette is used, it must be the 3.5 inch disk in MS-DOS format. It must be a new diskette and must bear very clearly the name(s) of the author(s) and the title of the paper. Authors must ensure that the version of the electronic copy is exactly the same as the paper copy that accompanies it. The software tools used must be Word for Windows or WordPerfect. Authors wishing to use any other software tools must first agree this with the Executive Editor. Neither the hard or electronic copies of manuscripts will be returned to the authors.

Submission fee: In line with the policy of providing a forum for dissemination of results of statistical research activities, no submission fees are charged for unsolicited contributions received.

The author: Each paper must carry the following information on the front page in this order: (1) the title (2) the name(s) of the author(s), (3) their institution(s)/affiliation(s), (4) a list of four or five keywords and (5) a short abstract of not more than 100 words. A clear indication of whom the proofs should be sent to (including the name, address, phone number, fax number e-mail address) should be given on this same page.

Format: Manuscripts should be printed on one side of the paper only. Pages should be numbered. All diagrams and graphs should be referred to in the paper as figures. Tables and figures are to be numbered in consecutive order in the text using Arabic numerals and should be printed on separate sheets.

References: References should be arranged in alphabetical order. Multiple references to the same author should be given in chronological order.

Footnotes: Footnotes should be kept to a minimum. When used, they should be numbered consecutively using Arabic numerals. Figures, tables and displayed formulae should not be included in footnotes.

Reproduction: Authors should note that printed copies will be made directly from photographic reproduction of final proof copies received from them. It is therefore imperative that high quality camera-ready originals are submitted. Illustrations should be of such quality that they are suitable for direct reproduction and ideally require the same degree of reduction. They should be clearly marked and correspond to references to them in the text.

Proofs: Two sets of proof copies will be sent to each author for final review. One of these must be signed and sent back to the executive editor within the time limit indicated in the cover letter.

Free copies: For each paper, author(s) will be entitled to one free copy of the journal of the issue in which the paper appears. The copy will be mailed directly to the author(s). Additional copies will be available at a special rate to the author.

Further information:

Enquiries relating to submission of papers etc. should be directed to:

Executive Editor

ROS, Eurostat, Room A2/162a

BECH Building

L-2920, LUXEMBOURG

Phone: +(352) 4301 34190 Fax: +(352) 4301 34149

e-mail: journal.ROS@cec.eu.int