ASC2005

MAXIMISING DATA VALUE

Data Use & Re-Use

ASSOCIATION for ASC SURVEY COMPUTING

An ASC International Conference on Survey Research Methods

Newland Park, Chalfont, Buckinghamshire, England, UK

Thursday & Friday, 15th & 16th September 2005

Under the joint organisation of :
Association for Survey Computing, Market Research Society,
Office for National Statistics, Royal Statistical Society

CONFERENCE PROCEEDINGS

ASC 2005
Maximising Data Value. Data Use & Re-Use

Proceedings of an

ASC Conference on Survey Research Methods

Newland Park, 15-16 September 2005

ORGANISED BY

The Association for Survey Computing (ASC)

The Market Research Society (MRS)

The Office for National Statistics (ONS)

The Royal Statistical Society (RSS)

EDITED BY

Raz Khan (ASC)
Cobalt Sky

Randy Banks (ASC)
University of Essex

Richard Cornelius (MRS)
ORC International

Suzanne Evans (RSS)
Birkbeck College

Tony Manners (ONS)
Office for National Statistcics

Association for Survey Computing

ASSOCIATION FOR SURVEY COMPUTING

PO Box 60, Chesham, Bucks, HP5 3QH, UK
tel/fax: +44 (0)1494 793 033
email: admin@asc.org.uk
http://www.asc.org.uk/

ISBN: 0 9546748 0 4

Preface

This is the third of the ASC's 2 day residential conferences focussing on methodologies in survey computing. These conferences aim to act as a bridge between our 3 day multi-stream international conferences which examine a broad range of issues and challenges, and our 1 day conferences which look at specific tasks or technologies. We felt that in our industry today (and in the world in general) there is pressure to increase efficiency and improve value for money, which leads us to look at uses for existing data and not "re-invent the wheel". We wanted to know what people were doing to make use of data already collected, and how they overcame problems of reconciling data from different sources. Computing facilities are getting more powerful, enabling us to design complex survey instruments and collect and process large volumes of data and we wanted to look at what might be happening at the leading edge. We also considered that respondent co-operation was diminishing, potentially endangering the whole foundation of our work. We wanted to know what steps colleagues are taking to maintain the goodwill of our interviewees in an increasingly "time-poor" society.

This led us to issue our call for papers in 2004, with the aim of exploring survey research methods in the area of data integration: making the most of existing data and metadata by using them as a platform to aid further research, using them for deeper secondary analysis, and combining multiple sources of data.

We assembled a programme committee, joining forces as in our previous methodology conferences, with MRS, ONS and RSS. We received many excellent abstracts and accepted the papers you find in this volume. They fell neatly into four broad themes, each of which was owned by one of the organisations:

Theme	Organiser
Process Integration	Tony Manners, ONS
Methodology and Software for Complex Models	Suzanne Evans, RSS
Models for Data, Metadata and Knowledge	Randy Banks, ASC
Multi-mode and Multi-Source Surveys	Richard Cornelius, MRS

Four eminent individuals were invited to give keynote addresses for each of these themes:

Karen Dunnell, recently appointed as the National Statistician and Director of ONS, discusses ONS's plans to redesign its continuous household surveys into a single survey. The contributed papers cover aspects of process integration and we are particularly grateful to Allyson Seyb and Philip Cookson who travelled from New Zealand and Australia respectively to address the conference.

Nicky Best of Imperial College leads the RSS's session and describes the use of Bayesian graphical models to combine multiple data sources. This session includes three papers looking in depth at the application of complex statistical models.

Andrew Westlake of SASC addresses issues of combining data and knowledge in models and illustrates them with experiences from a project looking to model simultaneously all forms of

passenger movement in London. The papers in this session look at bringing together disparate data in a variety of ways.

George Terhanian of HI Europe shares his extensive experiences of multi-mode research and data linkage. This session includes papers on disseminating data through web portals and enabling data re-use in the web-enabled world.

These proceedings follow the order of presentation of and within thematic sessions.

The organisers would like to thank all the speakers who have contributed significant amounts of their time in writing their papers and delivering them to this conference. We hope that you will find the event informative and stimulating, and that you will enjoy discussing the proceedings with your colleagues in the industry.

Scientific Programme Committee

Name	Organisation/Company	Affiliation
Randy Banks	ISER, University of Essex	ASC
Richard Cornelius	ORC International Ltd.	MRS
Suzanne Evans	Birkbeck College, University of London	RSS
Anthony Fielding	University of Birmingham	RSS
Paul Hewson	University of Plymouth	RSS
Raz Khan	Cobalt Sky Ltd.	ASC
Tim Macer	Meaning Ltd.	ASC
Tony Manners	ONS	ONS
Wendy Sykes	ISR	SRA
Andrew Westlake	SASC	ASC

Acknowledgements

This conference is the result of a huge team effort by the volunteer scientific programme committee listed above and we thank them all for the extensive time spent in selecting papers and preparing the programme. Special thanks go to Randy Banks, Richard Cornelius, Suzanne Evans and Tony Manners for the time spent in liaising with keynote speakers, and their thorough reviews of papers in their sessions.

There is a huge administrative effort in organising a conference and we thank Diana and Steve Elder for their work in managing the venue and bookings, Randy Banks for collating, formatting and printing the proceedings, and Hugh Gentleman for his artistry in preparing the publicity.

About *The Association for Survey Computing*

The Association for Survey Computing (ASC), originally known as the Study Group on Computers in Survey Analysis (SGCSA), was formed in 1971 in order to improve knowledge of good practice in survey computing and to disseminate information on techniques and survey software.

Today its aims are to:

- act as a forum for the various disciplines within survey research and statistical computing
- inform members of the latest software packages and techniques
- organise regular conferences and workshops on key topics within the industry
- disseminate information via its web site and publications
- catalogue current software systems

The ASC is a non-profit organisation, affiliated to the British Computer Society (BCS) and the International Association for Statistical Computing. (IASC) It has a wide-ranging membership at both individual and corporate levels, and has close working links with the Social Research Association, a well as the three co-organisers of this conference. Although based in the United Kingdom, it has a growing international membership.

The Association sponsors students so that they can attend courses in our discipline, and also sponsors prizes for well presented conference posters

The ASC organises four series of conferences:

- Three-day, multi-stream international conferences that cover all aspects of survey computing. The first one was in Bristol in 1992, then London in 1996 and Edinburgh in 1999, and the latest at Warwick in 2003.
- Two-day, single stream residential international conferences that have time to investigate the latest practices in survey computing. The first one was in Southampton in 1998, followed by Latimer in 2001; the current conference is the third in the series.
- One-day conferences (usually two each year) that concentrate on a single topic, methodology or medium.
- Occasional, specialised workshops that contrast and compare solutions as offered by different practitioners.

The ASC has recently joined forces with MRS to offer the 'MRS/ASC Award for Technological Effectiveness'. This award is intended to foster innovation in the involvement of computers in the survey process, whether on a large or small scale, by both companies and individual professionals. This award was presented for the first time in the autumn of 2003.

The activities of the ASC are organised by a committee of volunteers, supported by a part-time Administrator.

THE ASSOCIATION FOR SURVEY COMPUTING (ADMIN@ASC.ORG.UK)
PO BOX 60, CHESHAM
BUCKS HP5 3QH
TEL/FAX: +44 (0) 1494 793033
HTTP://WWW.ASC.ORG.UK/

About *The Market Research Society*

With members in more than 70 countries, MRS is the world's largest association representing providers and users of market, social, and opinion research, and business intelligence.

MRS serves both individuals and organisations who identify with its core values of professionalism, excellence, and effectiveness.

It has a diverse membership of individual researchers within agencies, independent consultancies, client-side organisations, the public sector and the academic community – at all levels of seniority and in all job functions.

MRS Company Partners include agencies, suppliers, and buyers of all types and sizes who are committed throughout their organisations to supporting the core MRS values.

All individual members and Company Partners agree to self-regulatory compliance with the MRS Code of Conduct. Extensive advice to support this commitment is provided by MRS through its Codeline service and by publication of a wide range of specialist guidelines on best practice.

MRS offers various qualifications and membership grades, as well as training and professional development resources to support them. It is the official awarding body in the UK for vocational qualifications in market research.

MRS is a major supplier of publications and information services, conferences and seminars, and many other meeting and networking opportunities for researchers.

MRS is "the voice of the profession" in its media relations and public affairs activities on behalf of professional research practitioners, and aims to achieve the most favourable climate of opinion and legislative environment for research.

THE MARKET RESEARCH SOCIETY (INFO@MRS.ORG.UK)
15 NORTHBURGH STREET
LONDON EC1V 0JR
TEL: +44 (0)20 7490 4911
FAX: +44 (0)20 7490 0608
HTTP://WWW.MRS.ORG.UK/

About *The Office for National Statistics*

The Office for National Statistics (ONS) is the government department that provides UK statistical and registration services.

ONS is responsible for producing a wide range of key economic and social statistics which are used by policy makers across government to create evidence-based policies and monitor performance against them.

The Office also builds and maintains data sources both for itself and for its business and research customers. It makes statistics available so that everyone can easily assess the state of the nation, the performance of government and their own position.

The Office also incorporates the General Register Office for England and Wales (GRO). The GRO is responsible for ensuring the registration of all births, marriages and deaths in England and Wales, and for maintaining a central archive dating back to 1837.

The National Statistician, Karen Dunnell, is the Director of ONS and Registrar General for England & Wales. The Office was formed in April 1996 when the Central Statistical Office merged with the Office for Population, Censuses and Surveys.

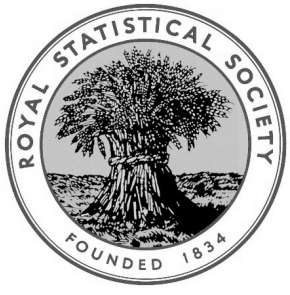
THE OFFICE FOR NATIONAL STATISTICS
1 DRUMMOND GATE
LONDON SW1V 2QQ
TEL: +44 (0) 845 601 3034
HTTP://WWW.STATISTICS.GOV.UK/

About *The Royal Statistical Society*

The Royal Statistical Society (RSS) is the UK's only professional and learned society devoted to the interests of statistics and statisticians. It is also one of the most influential and prestigious statistical societies in the world. The Society has an international membership, and is active in a wide range of areas both directly and indirectly pertaining to the study and application of statistics. The Society's activities include:

- Nurturing and stimulating the discipline and profession of statistics through publications and meetings
- Promoting the discipline and profession of statistics through its links with education, government, industry and the media
- Serving the interests of all statisticians through the setting of professional standards

THE ROYAL STATISTICAL SOCIETY (RSS@RSS.ORG.UK)
12 ERROL STREET
LONDON, EC1Y 8LX, UK
TEL: +44(0)20 7638 8998
FAX: +44(0)20 7614 3905
HTTP://WWW.RSS.ORG.UK/



Contents

MODELS FOR DATA, METADATA AND KNOWLEDGE

MULTI-MODE AND MULTI-SOURCE SURVEYS

Process Integration

Towards a Single Continuous Population Survey for the UK

Karen Dunnell

Abstract

This paper describes ONS' plans to redesign its existing continuous household surveys (GHS, EFS, LFS, Omnibus) into a single module-based survey. It covers rationale, methodology, efficiency and statistical benefits.

Keywords

Survey design, Continuous surveys, Household surveys

1 Introduction

This paper describes ONS' plans to redesign its existing continuous household surveys into a single module-based survey, known as the Continuous Population Survey (CPS). The surveys included for integration are:

- the Labour Force Survey (LFS) and associated boosts;
- the General Household Survey (GHS);
- the Expenditure and Food Survey (EFS), and;
- the Omnibus Survey (OMN).

Summary information regarding these surveys is provided in Annex A.

While these surveys individually have been successful, an integrated approach could deliver better value for money and increase the value of statistical outputs.

A comprehensive integration of the entire survey process is proposed; from the creation of a unified field force of interviewers administering a common modular questionnaire to the processing and production of outputs from a single source. As such, the CPS represents a move to a modular household survey system, is part of the ongoing programme of modernisation of the statistical infrastructure of the ONS, and is consistent with the aim of developing a world class framework for social statistics.

2 Rationale

At the heart of the Continuous Population Survey is a recognition of the need for National Statistics to produce better information on key social and economic variables between decennial censuses, for a

range of policy purposes, and to meet the increasing demand for regional and sub-regional information.

Demand for small area statistics has grown rapidly in the last decade, with particular pressure for information about income and ethnicity, but also other outputs too. There is also increasing demand for new surveys, including those under EU regulations which Member States are obliged to fulfil. And, there is an ongoing need to maximise value from ONS' continuous surveys and improve the coherence of National Statistics.

These demands cannot be met within the current survey arrangements. Individually, surveys have reached their limit in terms of length and burden and data are not easily pooled across surveys as their designs are different. In addition, the existence and maintenance of separate fieldforces, instruments, and processes represents a duplication of effort and sub-optimal use of limited resources.

The central objectives of the CPS are to:

- **develop a world class modular survey system**, better able to meet the information needs of the 21st century;
- **provide more coherent, better quality information** on which Government, stakeholders and the wider user community can base decisions;
- **create a range of new outputs from the core sample**, including inter-censal estimates of key socio-demographic variables at the sub-regional level;
- develop **a survey system with the flexibility to accommodate other surveys** at a later stage.

And, to achieve these objectives:

- while **maintaining continuity** of outputs and preserving the integrity of key time series;

without the need for resource additional to the combined budgets of the five component surveys, and;

- while **delivering further efficiency savings** from economies of scale and increased value of statistical outputs.

3 Design and methodology

Sample and fieldwork

The CPS will adopt an unclustered design, similar to that used by the Labour Force Survey in Great Britain. Currently, all other Government household surveys use clustered designs, where addresses in a sample are selected from particular areas grouped in small 'clusters'. Traditionally, this was necessary for all but the very largest surveys for reasons of economy. Clustering reduces interviewer travel time and costs, and makes fieldwork practical and affordable. However, a clustered sample leads to less reliable estimates.

By combining all the survey samples and fieldwork into one overall design the CPS, for the first time, will deliver unclustered samples for all interview combinations. This will allow sample size reductions for topics in the GHS, EFS and Omnibus Survey to achieve the same level of precision as now, or substantial precision gains if existing sample sizes are retained.

The CPS core sample is likely to be in the region of 250 thousand households and more than 500 thousand adults per annum, making it the largest ever continuous survey to be conducted in this country. However, annual samples are not wholly additive since there is some overlap because of the panel element within some interview combinations.

At present ONS has two separate field forces of interviewers conducting household surveys – one field force administers the LFS and associated boost surveys, while the other conducts all other household surveys. In preparation for the CPS, the data collection area will be restructured to deliver and support an integrated field force. This process involves major changes in organisational structure and procedures, changes in roles and responsibilities and more of a regional focus to fieldwork management and support. Under the CPS, each interviewer will be responsible for delivering all the required interview types in a small geographical area close to where they live, thus reducing travel time and cost and increasing productive contact and interviewing time to help maximise response.

The CPS will continue to use the Small User Postcode Address File to sample individuals living in private households and a small number of communal establishments. However, it could readily accommodate improved sampling frames or designs for expanding establishment coverage, should the opportunity arise.

Modular design

The CPS questionnaire will be designed as a single modular survey instrument comprising:

- a **core module** administered to the whole sample providing information on key variables for all CPS households and persons;
- **topic modules** administered to parts of the sample providing information on variables for which sufficient precision to meet policy needs can be obtained from a portion of the CPS sample;
- a small number of viable **interview combinations** formed from combining core with selected topic modules so that all topic modules are covered.

The project involves the design of a new core module and a re-examination of the combinations of topic modules represented by the current surveys, with the aim of improving coherence in reporting and optimising the use of modules. Decisions on the exact content of the core module and the restructuring of topic modules will be made as part of the ongoing consultation process with stakeholders.

Core module

The core questionnaire needs to be relatively short and straightforward so that total interview length for core and topic modules remains viable. Indeed, it is essential to the success of the CPS that the core questionnaire does not become unduly burdensome. Questions that are eventually included in the CPS core are likely to meet all or some of the following criteria:

- a classificatory variable essential for analysis
- an output for which there is a clear requirement for a high level of precision nationally or regionally, and not provided elsewhere
- an output for which there is a clear requirement for reporting at a sub-regional level, for example to local authority or health authority district level, and not provided elsewhere
- question(s) that can be administered by either face-to-face or telephone interviewing, and for which proxy responses are acceptable
- question(s) which would not adversely affect response to the CPS as a whole

There would be advantages to fixing key core questions over a number of years to provide an uninterrupted time series to a high level of precision. However, continuous reporting is not essential for all outputs. To allow greater flexibility and scope for inclusion of a wider range of questions, it

would be desirable to include some questions periodically. Therefore, it is proposed to sub-divide the core module into two modules:

- A **fixed core module** with questions normally included for at least a 5 year period, or longer
- A small **rotating core module** with questions normally included once every 3 years. For example, where a question might be asked in 2008, 2011 and so on.

Topic modules

Questions in topic modules will collect detailed information on the various subjects covered by the existing surveys where a continuing need for this information is identified, for example, health, education, labour market, income and expenditure.

It is possible to run some modules across more than one interview combination to generate a sufficient sample size. It would also be possible to use the core questionnaire to identify rare or hard-to-find groups so that additional questions could be administered, or respondents followed-up at a later stage. A modular design would enable the quicker and more effective introduction of new modules, and amendment of existing modules, as customer output requirements change. The diagram in Annex B illustrates how the CPS might be organised to deliver a range of existing and new outputs.

Survey procedures

The CPS will be a mixed-mode data collection instrument. Where an interviewer records an initial non-response, households may be 're-issued' to an alternative interview mode for a further attempt to secure an interview where appropriate. For example, a field interview may be re-issued to the telephone unit where a telephone number can be obtained, or a telephone interview re-issued to a field interviewer.

An integrated field force, combined with updated data handling systems will improve the flexibility and speed with which re-issuing of interviews between modes can occur. An efficient and systematic approach to mixed mode interviewing will help maximise response.

Different topic modules may have different rules for the use of personal, telephone or self-interview, as with the current surveys. However, it is proposed that all questions in the core module should be suitable for both personal and telephone interviewing modes. Therefore, at a minimum, in the event of interview non-response in the field, the core module could be re-issued to the telephone unit.

In addition, electronic questionnaires will be redesigned to provide:

- a **standardised** approach to asking questions. dependent on a respondent's answer from a previous interview (known as 'dependent interviewing');
- improved **flexibility** to allow individuals in a household to be interviewed consecutively, concurrently, or in combinations of both ('concurrent interviewing');
- **intelligent questionnaires** capable of recognising the mode of interview and adapting accordingly, and;
- more extensive use of **context sensitive help** to resolve interviewer and respondent queries as they arise

4 Process integration

While there is already much standardisation in the way ONS household surveys are designed, collected, processed, analysed and disseminated, there is considerable scope for further streamlining

and standardisation with the CPS, particularly as part of the wider modernisation of ONS's statistical and computing infrastructure. The transition from a complex portfolio of surveys to a streamlined household survey system is illustrated in the diagrams below, where process 1 might represent the organisation and training of interviewers, process 2 the design and administration of questionnaires, and process 3 the editing, imputation and weighting of data etc.

Figure 1 shows an illustration of the extensive system of separate, but related, processes involved in each survey.

Figure 1 : Process model of existing surveys

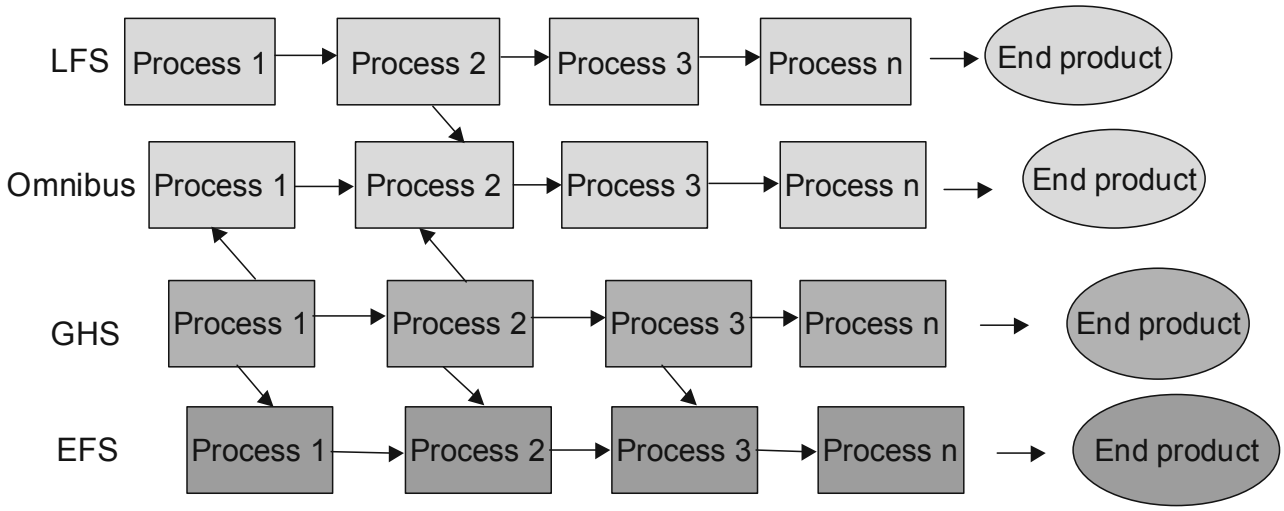
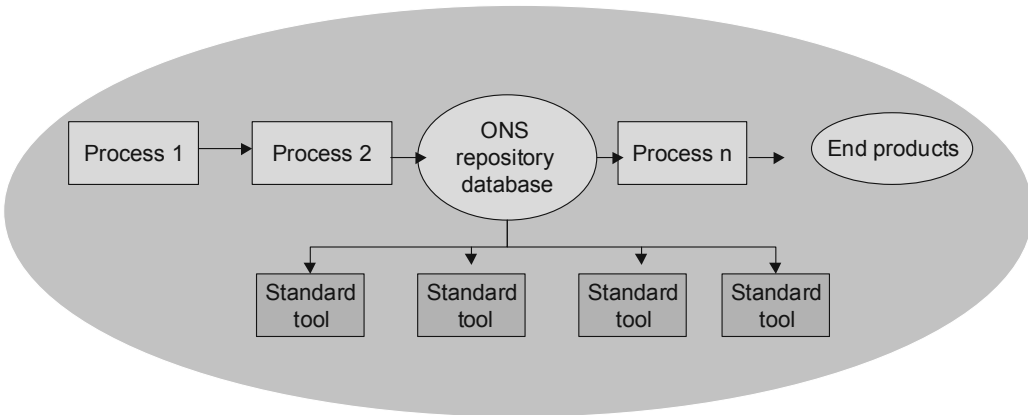


Figure 2 illustrates the vision for how the above model will be replaced with a single and efficient production chain. Data are derived from one questionnaire, administered by a single field force, and stored in a central repository database (CORD). Editing, imputation and weighting are carried out using a set of common statistical tools. A range of end products are still derived, but from a common source.

Figure 2 : Process model for Continuous Population Survey



A set of standard statistical tools and the central repository database mentioned above are being developed alongside the CPS as part of the statistical modernisation programme. In addition, improvements are being made to the data handling and field management system in order to handle the flow of survey information and manage field operations more effectively.

5 Statistical benefits

A single modular survey approach will improve coherence in official statistics by delivering a range of outputs from a single data source, reducing 'competing estimates' between surveys. The process of harmonisation is furthered since core module questions will be asked in the same order across all interview combinations and will be, by definition, common across the survey.

The very large sample of core data will provide better quality and more reliable estimates on key social and economic variables between decennial censuses. It will also help to meet the increasing demand for regional and sub-regional information that cannot currently be met by the existing surveys. A range of new outputs will become available. For example, banded household income will be collected on the core module and together with information on topics such as housing, health, employment and education will provide regular information down to local authority district level.

The flexibility of the modular design will enable ONS to meet the requirements for new surveys more readily and cost-effectively than at present. It also provides the flexibility to plan topic modules on the basis of sample size for the level of precision required, rather than constrained by whichever existing survey vehicle happened to be available, as at present.

As well as developing an appropriate design for the planned 2008 launch, the new survey is being built with future change in mind. The expectation is that new modules will be added as policy needs change over time, and that requirements for some new surveys will be met through the CPS. In addition, sponsors of other surveys may opt to switch their existing continuous surveys to this vehicle. Each survey joining the CPS would contribute to the overall sample size and further boost the precision and analytical power of the survey data.

Therefore, an essential element of CPS development is to 'future-proof' the survey. The sample structure will be designed so that wholesale changes will not be required, even if sample size requirements alter substantially. The modular structure of the survey instrument will readily accommodate new topics, while a range of survey types and features can be incorporated within the CPS survey system. For example:

- interview combinations with cross-sectional, quarterly or annual panel designs
- diary components and other self-completion elements
- telephone or personal interviewing
- interviews with all household members or with individuals sub-sampled within a household, or combinations of both

Hence, the data collection method chosen for each topic module or interview combination can be based on statistical requirements rather than the constraints of a particular survey vehicle.

6 Efficiencies

Moving to an integrated survey will deliver better value for money from the ONS household survey programme. It will also help to reduce risks associated with our existing systems and deliver efficiencies through greater standardisation, integration, modernisation and better design.

By repackaging its surveys into a single modular design, ONS will increase the value of data already collected and deliver new, regular outputs from the large core module, without any increase in costs.

There will be economies of scale, for example in survey management costs, quality assurance, and development costs associated with moving from five surveys to one. There will also be reduced field management costs from integrating field forces, reorganising data collection functions and through introducing modern survey management systems. Currently, inadequate management information systems hinder the management of surveys, including allocation of work, the real-time monitoring of interviewer performance and tracking response rates across ONS surveys.

Some of the existing survey systems are creaking and present a growing risk to timely outputs. More efficient, effective, flexible and robust systems will provide a wider range of functionality and outputs than is currently available. In addition, a fully integrated survey system will be more reliable, and easier to maintain and update than under current arrangements which should lead to reduced IT support costs, through the need to maintain and support only one survey system.

Furthermore, new continuous household survey work will frequently be able to be done cheaper, and sometimes much cheaper, through better. A flexible survey infrastructure will exist that can be adapted to take on new demands. It will be straightforward to add new topic modules or a new interview combination to the overall design.

7 Development timetable

ONS plans to launch the CPS in January 2008. Before this, a series of trials will test, develop, and validate all aspects of the proposed survey. These started in March 2005 with a small scale feasibility test using experienced interviewers. A second feasibility test is planned for September 2005 which will build on lessons learnt from trial 1 and provides the opportunity to test out the new survey on less experienced interviewers.

Work is presently ongoing on a range of methodological issues concerning question development and data collection. Combinations of qualitative and quantitative techniques are being employed to investigate new questions, relating to income and education, for example, and to asses output quality and comparability. A range of issues will be examined, including item and unit non-response, mode of data collection, collection of proxy responses, order effects, and questionnaire flow. In each area, the nature of change will be established, and its effect understood and minimised.

Fieldwork modernisation and integration will be phased in gradually over the next 12-18 months in readiness for the CPS launch. The new style of interviewer work package and training and, in particular, the potential impact on response will be tested during a large scale pilot in spring 2006. A second pilot will be conducted towards the end of the year to quality assure necessary improvements to the survey case management system on which the CPS is heavily dependent for successful deliver.

CPS development timetable

2004	Initiate project – build up team, develop plans
July 2004	Publication of initial consultation document
Autumn 2004 – spring 2005	Develop modular design
Spring & summer 2005	Small scale field trials; further work on questionnaire design issues
2006 – mid 2007	Fieldwork modernisation, field force integration
	Large scale pilot fieldwork
Second half 2007	Transition period; survey training, survey management reorganisation
January 2008	CPS launched
January – March 2008	benchmark CPS and existing survey estimates for continuity purposes

In parallel with the first quarter of CPS data collection, ONS plans to continue to collect some survey data using current survey methods and design, in order to benchmark outputs for continuity purposes; the LFS will be the top priority. The design of the parallel run has yet to be agreed but its purpose will be to enable ONS to use the data to rework back series so that a new consistent time series can be produced in the event of serious discontinuities arising.

Appendix A. Surveys for integration: the LFS, GHS, EFS and Omnibus survey

The General Household Survey (GHS), the first multi-purpose household survey, started in 1971 and covers a wide range of social and socio-economic topics. The main aim of the survey is to collect data on core topics including housing, employment, education, health and family information.

The Expenditure and Food Survey (EFS) started in 2001 bringing together two surveys, the Family Expenditure Survey (FES) and National Food Survey (NFS), that were both well established and important sources of information, charting changes and patterns in Britain's spending and food consumption since the 1950s.

A Labour Force Survey (LFS) has been carried out in the UK since 1973 and in its present form since Spring 1992, providing a wide range of data on labour-market statistics and related topics such as training, qualifications, income and disability. In recent years the quarterly LFS has been supplemented by a series of annual boost samples in England, Wales and Scotland.

The National Statistics Omnibus Survey (OMN) is a regular, multi-purpose survey that started up in 1990 in order to provide quick answers to questions of immediate interest and information on topics that do not require a full, in-depth survey.

All the surveys attempt to conduct interviews with every household member, with the exception of the Omnibus survey where interviews are administered to one randomly selected household member only.

The surveys differ in their use of Computer Assisted Personal and Telephone Interview (CAPI and CATI) and the extent to which proxy responses are permitted.

The table below summarises some of the main features of the current surveys.

Current data collection features of the four component surveys						
	CAPI [Computer Assisted Personal Interview]	CATI [Computer Assisted Telephone Interview]	Diary	Proportion of proxy responses	Who interviewed?	Panel component?
EFS	✓	✗	✓	14%	All >16	✗
GHS	✓	✓	✗	5%	All >16	✗[2]
LFS	✓	✓	✗	32%	All >16[1]	✓
OMN	✓	✓	✗	Varies	Selected adult	✗

[1]Economically inactive respondents over the age of 70 are not interviewed in waves two to five of the quarterly LFS, other than to check their economic status is unchanged.

[2]The GHS will incorporate a panel component from 2006.

More information about these surveys can be found via the following links.

Labour Force Survey
http://www.statistics.gov.uk/StatBase/Source.asp?vlnk=358&More=Y#general

Labour Force Survey boosts
http://www.statistics.gov.uk/CCI/article.asp?ID=370&Pos=&ColRank=2&Rank=416

Expenditure and Food Survey
http://www.statistics.gov.uk/ssd/surveys/expenditure_food_survey.asp

General Household Survey
http://www.statistics.gov.uk/StatBase/Source.asp?vlnk=263&More=Y

National Statistics Omnibus Survey
http://www.statistics.gov.uk/services/SurveyOmnibus.asp

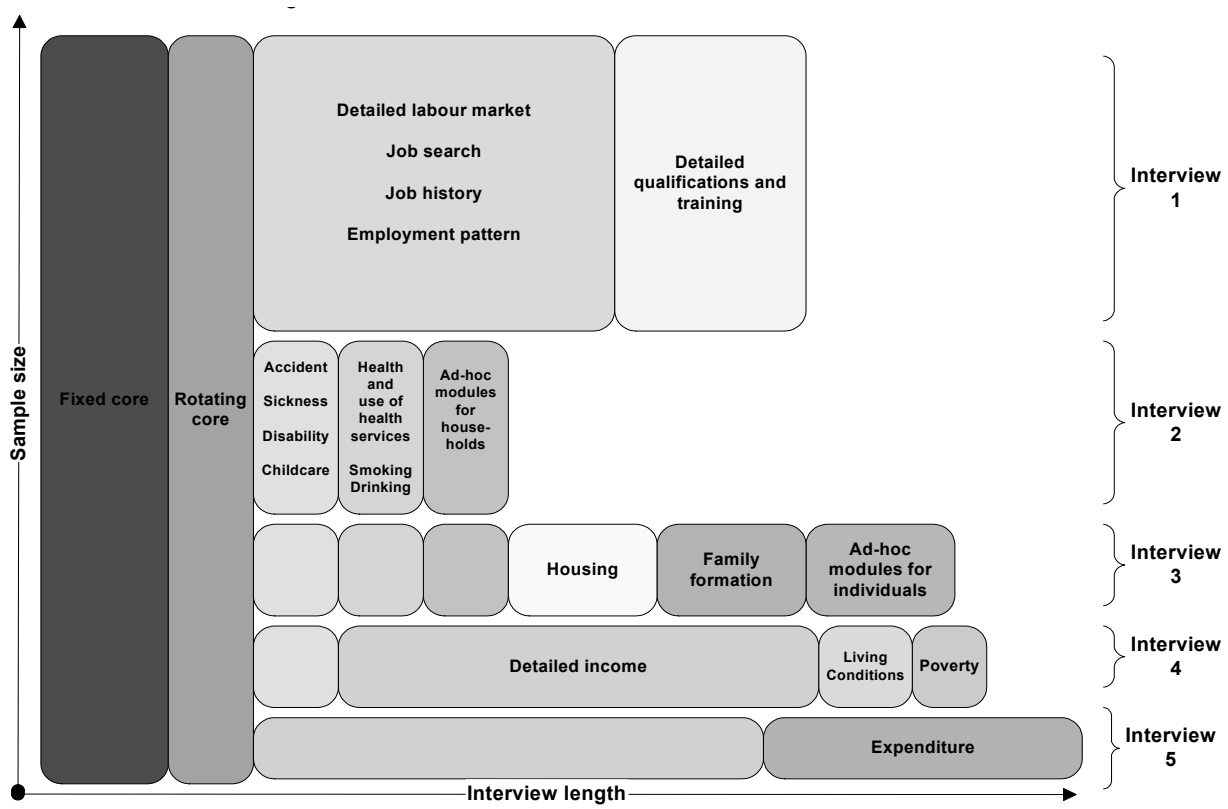
Links to other documents

Proposals for the Continuous Population Survey
http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9668&Pos=&ColRank=1&Rank=272

Neighbourhood Statistics Service (NeSS)
http://www.neighbourhood.statistics.gov.uk

ONS Modernisation
http://www.statistics.gov.uk/about_ns/BusinessPlan/Business_Plan_2001-04/modernising.asp

Appendix B. Illustrative Diagram of a Modular CPS Questionnaire



Notes

Interview 1 represents an LFS type interview combination. It preserves the quarterly and annual panel designs within the current LFS arrangements.

Interview 2 represents another viable set of interview topics for a single interview.

Interview combinations 3 and 4 represent a range of outputs currently collected by the GHS and Omnibus surveys, and, in future by the prospective Survey of Income and Living Conditions. It is particularly difficult to foresee the development paths of these surveys over the next few years. Plans will need to be adapted accordingly as these become clearer. In this model, interview type 3 is cross-sectional, while interview 4 is based on a annual panel design.

Interview type 5 represents an EFS type interview, cross-sectional and with a diary component.

About the Author

Karen Dunnell is the National Statistician and Director of the Office for National Statistics. She began her career as a health care researcher and joined the Office of Population Censuses and Surveys (OPCS) in the mid 1970s carrying out a wide range of national surveys and managing medical statistics. When OPCS merged with the Central Statistical Office to form the ONS she became the Director of Demography and Health Statistics. She later moved to a central post to help launch National Statistics and plan the arrival of Len Cook, the first National Statistician in 2000. She then became a Group Director in Social Statistics in 2000 and became an ONS Executive Director in 2002. She was responsible for setting up the new "Sources" Directorate, bringing together household and business surveys, the infrastructure that supports them and planning for the 2011 census. She also launched the ONS Statistical Modernisation Programme.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Statistics New Zealand's Longitudinal Business Frame

Allyson Seyb

Abstract

This paper introduces Statistics New Zealand's new longitudinal research database, the Longitudinal Business Frame (LBF). Two different techniques, one based on probabilistic matching of firm births to existing firms, and the other based on tracking changes in employer administrative numbers by tracking their employees, are used to repair broken longitudinal links in the LBF. The LBF, with its economy-wide coverage and basic attributes such as employment, location, industrial activity and ownership relationships, is a rich source of longitudinally linked business data.

Keywords

Longitudinal data; record linkage; administrative data

1 Introduction

The development of a longitudinally linked business database, the Longitudinal Business Frame (LBF), is part of a strategic research project at Statistics New Zealand. The Linked Employer Employee Data (LEED) project integrates existing employer and employee information to provide new insights into the operation of the New Zealand labour market and its relationship to business performance. LEED combines administrative data drawn from the New Zealand taxation system, with business data drawn from Statistics NZ's Business Frame (BF). One of the key advantages of the LEED data over existing data is the ability to follow businesses and individuals over time.

Statistics NZ's BF provides current, point-in-time snapshots of business populations; at any time the BF contains the most recent information Statistics NZ has about a business. The BF is limited to those firms deemed to be 'economically significant' according to the amount of Goods and Services tax (GST) they pay. Changes on the BF are recorded attribute by attribute in 'history' tables. To identify an enterprise's attributes at an earlier point in time, it is necessary to undo changes applied to the enterprise since that time. The BF history tables are used to create a longitudinal series that forms the basis of the LBF. Tax data from NZ's Inland Revenue Department (IRD) is used to extend the coverage of the LBF to include small businesses. The LBF provides longitudinally linked business data for all businesses in the LEED database, including businesses that are too small to be in the BF. The LBF contains information on business activity at both the plant ('establishment') level, and the enterprise level, from April 1999.

One of the main issues associated with longitudinal business data is handling changes in structure. The LBF attempts to overcome this by providing information about mergers, acquisitions and splits

observed in the database. This enables researchers to identify changes in structure and treat them accordingly.

Research using annual snapshots of the BF found problems with spurious births and deaths due to broken longitudinal linkages (Carroll et al, 2002). Sometimes, businesses change administrative numbers within the tax and BF systems without undergoing any change in the real world, for example, when a business 're-brands'. If the identifiers have changed then the longitudinal links are broken. Two different methods are used in the LBF to identify changes in administrative identifiers in the source data: the first depends on probabilistic matching techniques applied to BF births, and the second involves tracking groups of employees that change employer IRD number en masse.

The LBF has many potential uses: an enhanced measure of business demographic data; a source of age information for businesses; identifying and tracking the components of firms involved in structural change; and linking to other business surveys, such as the Annual Enterprise survey, allowing the use of business survey data longitudinally for productivity and performance measurement. Other international economy-wide longitudinal databases include the US Census Bureau's Longitudinal Business Database (Jarmin and Miranda, 2002) and the UK Office for National Statistics' Annual Respondents Database (Barnes and Martin, 2002).

Results reported in this paper are based in part on tax data supplied by the IRD to Statistics NZ under the Tax Administration Act. This tax data must be used for statistical purposes only and no individual information is provided back to the IRD for administrative or regulatory purposes. Careful consideration has been given to the privacy, security and confidentiality issues associated with using tax data in this project. The IRD collects this data to support the efficient operation of the New Zealand taxation system. Its use as a base for the production of statistics places new and quite different demands on the data. Any discussion of data limitations or weaknesses is in the context of this latter use, and is not related to the ability of the data to support the IRD's core operational requirements.

This paper describes the creation of the LBF. Section 2 describes the source data and discusses the two main issues that apply to longitudinal business data: businesses that change structure or administrative numbers over time. Section 3 outlines the use of probabilistic matching to repair broken longitudinal links between businesses. Section 3 also looks at an alternative source of information about changes in administrative numbers – employees. In section 4, we discuss some features of the LBF and section 5 concludes and discusses next steps.

2 Source data for the LBF

The LBF contains information from two main sources: the Statistics NZ Business Frame and tax records drawn from the New Zealand taxation system.

The Business Frame

Statistics NZ has maintained a business frame since 1986. The Business Frame is a database of individual, private and public sector businesses operating in New Zealand. The information recorded about each business on the BF includes: name; IRD number; location; ownership links with other businesses; numbers of persons employed; and codes indicating industry, institutional sector, type of business and the level of overseas ownership. The BF is used as the frame for business surveys and as the source data for business demographic statistics.

The BF uses a three level statistical unit model comprising the enterprise, the kind-of-activity unit and the geographic unit. The individual components of the unit model are as defined in International Standard Industrial Classification of all Economic Activities, Third Revision (ISIC rev 3). The primary unit on the BF is the enterprise and, in general, it is the enterprise that corresponds to a tax reporting unit. The enterprise represents a legal entity, which may be a company, partnership, trust, estate, incorporated society, producer board, local or central government organisation, religious organisation, voluntary organisation or self-employed individual. The kind-of-activity unit is an enterprise or part of an enterprise that engages in predominantly one kind of economic activity without being restricted to a geographic area. The geographic unit is defined as a separate operating unit engaged in one kind of economic activity from a single physical location. A Statistics NZ 'geographic unit' is equivalent to a 'plant' as defined in the economic literature and to an 'establishment' as defined in the System of National Accounts 1993.

Most commonly, enterprises have only a single kind-of-activity unit and a single geographic unit. In this case, the enterprise does business at only a single location and all three units refer to the same economic entity. Alternatively, an enterprise may have several geographic units and one or more kind of activity units. Generally, businesses on the BF are uniquely identified by their enterprise numbers.

Coverage of the BF

While the target population of the BF is all businesses and organisations engaged in the production of goods and services in NZ, the actual BF population consists of all businesses over certain minimum thresholds. To be in scope for the BF the business must be registered with the IRD and fulfil any of the following conditions: have greater than $30,000 annual GST turnover (or greater than $40,000 IR10 income if the business is not registered for GST); have paid employees; be affiliated with other businesses; or be operating in the agriculture or forestry industries. The unreported economy is considered out of scope of the BF. The BF is designed to provide an up-to-date and accurately classified register of economically significant NZ businesses; at March 2005, the BF listed approximately 420,000 active enterprises.

The BF is updated continuously with administrative data from other government agencies, as well as data collected by Statistics NZ. In fact, the frame relies heavily on the IRD's Client Registration database, a database comprising all taxpayers, excluding individual wage and salary earners, and is updated monthly by detecting changes in the Client Registration database. In addition, new IRD registrations which meet any of the criteria listed above are added to the frame monthly.

Data accuracy on the BF varies with the size of the business. Maintenance is targeted towards large and complex businesses, currently defined as greater than $200,000 annual turnover; BF fields are updated at least once every three years for these businesses and more often if changes in activity warrant it. Small businesses are added to the BF, and maintained thereafter, solely from IRD information. Businesses that de-register for tax or show no tax activity for a period of 12 months are ceased on the BF. Again, in keeping with the targeting of resources to larger units, deaths of larger units are confirmed before action is taken, while smaller units are ceased automatically.

Changes over time

Apparent changes in LBF outputs over time may be real or caused by administrative changes in the source data or in the LBF itself. The BF has undergone significant changes since 1986, including a migration from a centrally-controlled mainframe system to a Local Area Network (LAN) system in 1996, and a major change in industry coding, from ANZSIC96 version 1 to version 4 in March 1999.

The focus of this paper is on characteristics of the BF since April 1999, the first month of the LBF series.

There have been several key changes since April 1999. The BF population has changed over time. Prior to 2003, the population was predominantly defined in terms of 'economic significance' as described previously and there were well-documented areas of under coverage. Since then, greater use of existing tax data for frame maintenance has resulted in improvements in BF coverage, timeliness and accuracy. The coverage of the BF has been extended to include all employing businesses and all known areas of under coverage, particularly in those industries in which most of the activity is GST exempt, have been addressed.

The key fields on the BF from the point of view of the LBF are the reference numbers (both statistical unit and tax reporting unit), industry, region and employment fields. Identifiers on the BF are well defined and their application has not changed over time. However, from June 2003, more effort has been made to identify businesses that change legal structure, by specifically asking about this in the monthly birth questionnaire that is sent to all of the medium and large births. Traditionally, industry was collected via questionnaire at birth and thereafter updated annually. Since 2003 most BF births have been coded electronically using industry information from the IRD. Employment information on the BF changed from respondent-sourced full-time-equivalent numbers of employees to a predominantly tax-sourced head count of employees in 2002. Industry codes for enterprises are derived from the codes on the geographic units based on the number of employees carrying out each activity; when the source of the employment information changed there were a number of enterprises that changed industry as a result. For large businesses, fields such as employment, location and industry are checked annually via questionnaire. Small businesses are updated automatically from tax data as necessary, and may never be contacted by Statistics NZ.

LEED tax data

Income derived from salary and wages has PAYE (pay-as-you-earn) tax deducted by the employer. All employers must file an Employer Monthly Schedule (EMS) monthly to the IRD. Large employers file twice a month. Every month the employer lists each employee and the employee's details. The LEED project links EMS employer-employee information to the LBF; LEED uses the EMS data as a base and adds industry and region information from the LBF. In addition, information from the EMS data is used to repair longitudinal links between employers, in both the LEED system and in the LBF system.

Employees and employers are uniquely identified within the tax system by their IRD numbers. An EMS record represents an employee employed by a particular employer in a given month, plus employee name, tax and earnings details. This pairing of employers and employees means it is possible to identify administrative changes by following groups of employees that change employer IRD number in consecutive months.

EMS data, in its present form, has been collected by the IRD since April 1999, and collection and maintenance practices, and population exclusions and inclusions, have not changed substantially during that time. Statistics NZ receives EMS data from IRD monthly; the largest tables hold approximately 200 million employment records (March 2005) and are growing at the rate of 3 million records each month.

Births, deaths and restructures

In order to produce meaningful statistics about firm performance and dynamics it is necessary to be able to link businesses over time. A firm may undergo several changes in its lifetime, in addition to birth and death. For example, legal or administrative entities closing down or being created due to break-ups, mergers, split-offs, takeovers, or re-structuring. It is important that the true creation or destruction of firms, as opposed to administrative reshuffles, can be identified in the LBF system.

The Eurostat Business Register Recommendations Manual – March 2003 Revision defines birth and death as follows: *"a birth amounts to the creation of a combination of production factors and a death amounts to the dissolution of a combination of production factors, both with the restriction that no other enterprise is involved"* and goes on to say that *"an enterprise is considered to be continued if its production factors are continued. It is discontinued if its production factors are discontinued"*. It also states that *"A distinction must be made between the (conceptual idea of) real observable world and its reflection in administrative files and in statistical business registers"*.

Firm births are recorded in the LBF source data as follows: a birth appears as a new statistical unit on the BF; a birth appears as an employer IRD number used for the first time. Similarly for firm deaths: a death appears as a ceased statistical unit on the BF; a death appears as an IRD number that ceases to be used. In both of these data sources, businesses (that is enterprises on the BF and reporting units in the tax system) can change their identifier and appear as a birth or death within those systems. These are referred to as administrative changes, as they occur in the absence of a birth or death in the real world.

For example, an existing firm may begin to use a new IRD number for administrative convenience or because of a change in legal character. The firm will begin using the new IRD number to file its EMS form and, if the firm is large enough, the 'new' IRD registration will appear on the BF as a birth. There are several possible scenarios including:

1. An existing firm A assumes a new legal status as firm B with a new tax registration, for example, changing from an individual to a partnership. The firm continues to use the existing factors of production.
2. Existing firm A ceases. Some or all of its functions continue as new firm B.
3. An existing firm A splits off some of its functions to a new firm B. A continues trading.
4. Existing firms A and B amalgamate to form a new firm C.
5. Existing firm A acquires all of the production factors of firm B.

While all of the situations described above involve the birth of new administrative numbers, none of them is regarded as the birth of a new firm; these are all cases of *administrative churn*. Firms cannot be linked over time by enterprise number or IRD number, because either new numbers have begun to be used or existing numbers have ceased to be used. Administrative churn peaks at the end of one tax year and the beginning of the next; in New Zealand, where the most common balance date is at the end of March, this corresponds to a peak in April (Kelly, 2003).

Where reference numbers change in the absence of corresponding real world events, the ability to track continuing firms over time is lost. LEED and the LBF include processes that recognize when firms in different periods are the same business, even when their identifiers change in the source data. These processes are said to 'repair' the respective IRD and LBF identifiers. Both data sources contain information on changes in economic entities which can be used to infer the existence of real world events and undo spurious births and deaths: the longitudinal links between businesses established in the building of the LBF and the longitudinal links established between employers in the tax system.

3 Repairing longitudinal links

Longitudinal links in the LBF are broken when a business changes its unique identifier within either (or both) of the administrative databases from which the data is drawn. This is a fairly common occurrence: businesses that 're-brand' or are bought or sold receive a new IRD number in the tax system; new tax registrations appear on the BF as new enterprise structures with new reference numbers. It is important that these spurious births and deaths are identified and LBF records that belong to the same entity are linked. Failure to identify this link will affect our ability to produce statistics on the dynamics of new firms as opposed to continuing enterprises.

Changes on the BF are recorded attribute by attribute in 'history' tables. When changes are made to an attribute (e.g. industry code), both the date of the change on the BF and the date of the change in the real world are recorded. In this way, the history tables can be used to create monthly snapshots of the BF. Two known exceptions to the recording of changes for attributes within the history tables relate to large-scale automated updates and changes to IRD numbers. The first case includes major revisions of industry codes and annual geographic reclassification of frame units.

Seyb (2003) developed a methodology for creating a longitudinal business series based on these monthly snapshots of the BF. The longitudinal series created is augmented by the addition of small businesses that are present in the EMS data but not present on the BF. These 'IRD-only' businesses have limited industry information and no region classifications. The original methodology has been extended by the inclusion of processes to repair the longitudinal links.

Repairing longitudinal links using probabilistic matching techniques

The term 'record linkage' refers to "the bringing together of information from two different records that are believed to belong to the same person, family or entity. The records to be compared may come from a single data file or multiple data files."(Gill, 2001). There is a large amount of literature available which discusses the methodology of record linkage (e.g. Gill, 2001; Winkler, 2001). The two most commonly used methods for record linkage are the exact, or deterministic method, and the probabilistic method. Exact methods rely on the exact comparison of common variables – records either match, if the comparison agrees exactly, or don't match. The main requirement for exact matching is the presence of a variable that is universally available, fixed, easily recorded, unique to that entity and readily verifiable (Gill, 2001, Section 3.1, p27). Probabilistic methods use a combination of partial identifiers, for example names and addresses, which are used to compute 'weights' based on probabilities for each potential match. The resulting weight is then compared to a threshold to decide whether to mark this pair a truly matching pair (Gill, 2001). The addition of weights assumes the absence of a correlation between errors in the various items of identifying information (Newcombe and Kennedy, 1962). So, for example, if we are using name and address as matching variables we assume that the errors in names are independent of the errors in addresses.

If all possible combinations of record pairs are tested against each other, the number of comparisons would be very large. Rather than compare all record pairs, in practice it is common to consider only those pairs that agree on certain identifiers or *blocking criteria.* Variables not used for blocking can be used for matching. A set of blocking and matching variables is termed a *pass.* An efficient matching strategy involves using multiple passes, each pass independent of the previous one. While the majority of links that result will be true matches, there are usually a small proportion of *false positives* (records that have been linked but that in reality are not the same business). In addition, a few true matches will

be missed (*false negatives)*. Generally there is a trade-off between the two types of errors since, for example, reducing the rate of false positives may increase the rate of false negatives (Bycroft, 2003).

Ideally, businesses on the BF are linked over time by their geographic numbers. When a business changes administrative numbers, and this change has been detected by Statistics NZ, the original geographic unit is transferred to the 'new' enterprise and the new geographic unit invalidated. In this way, the longitudinal dimension is maintained. In the historical BF data many of the longitudinal links between businesses have already been identified through the usual BF maintenance processes. That is, geographic units have already been transferred between enterprises. Initial investigations showed that approximately 80 percent of the changes in administrative numbers identified on the BF are found within three years of the date of the change in the real world.

The repair of the LBF longitudinal links is composed of two main components. The first component identifies geographic units that have been transferred on the BF, these are units where it is unnecessary to carry out matching as the link has already been identified on the BF. Large births, for example, sent a questionnaire to collect their details are asked if the birth is a result of a change of ownership or change in legal structure. The second component involves probabilistic matching of records in two files: a file of births to the BF in a particular month and a file of businesses that were on the LBF in the previous month. The tool used is Ascential's QualityStage v7.0.1 (formerly Integrity). Geographic units are matched on variables common to both files, including name (both trading name and standardized legal name), location (standardized street address and region codes), industry code and phone number. This mix of variables contains some that are highly discriminatory (e.g. phone number, legal name) and others that are less so (e.g. street address and other location information). The matching routine processes the data in six passes. The first two passes limit potential matches to units in the same area (*meshblock*) and look for an exact match on legal or trading name. The third pass takes advantage of the discriminating power of the telephone number and uses this as a blocking field together with a combination of location, industry and name matching information to identify continuing businesses. Later passes cast the net more widely by gradually extending the size of the region that units to be compared may belong to and using various combinations of location, industry, and name matching fields, and increasingly higher cut-off thresholds, to identify links where one or more of the fields is in error.

A linked record pair occurs when we bring together two records, each with a different BF administrative number, but belonging to the same business. When a link is made the LBF is updated by recording the details of the new identifiers. For details of the repair of longitudinal links in the LBF using probabilistic matching techniques, see Seyb (2004).

Clerical review showed that the level of false positives was approximately 2-3 percent in any month, while the level of false negatives was no more than 5 percent. Using this method, approximately 15 percent of firm births in any month are shown to be continuing businesses. At the beginning of a tax year (April), the percentage of spurious births can be as high as 20 percent. This confirms the importance of the use of probabilistic matching to enhance the value of the LBF data. Our aim is to identify as many continuing businesses as possible, without ending up with too many erroneous links. Missing matches inflate the number of firm births and deaths.

Repairing longitudinal links by tracking employees

The EMS payroll return is filed monthly and covers all payers and recipients of income that is taxed at source, other than interest and dividends. Two types of recipients of income are covered by the EMS:

those who pay PAYE income tax, and those who have withholding payments deducted. Generally, those individuals who have PAYE deducted are employees, while those who pay withholding tax are a subset of the self-employed (Kelly, 2003). Employers report their tax using the same unique identifier, the employer IRD number, each month. Employers can be linked over time using the employer IRD number. Spurious birth and death of reporting units in the EMS data occurs when employees cease to be reported under a given employer IRD number, but in reality continue to work at the same location. For example, Graham Smith, a plumber, re-brands as Smith Plumbing Services Limited. He is re-registered with the IRD and receives a new employer IRD number for his 'new' business. Next month he files his EMS return using his new reference number. Links can be established with a high degree of certainty between the predecessor and successor employer IRD numbers by tracking Graham's employees within the LEED system.

The method is effective when predecessor firm size is greater than three employees, and at least 70 percent of employees 'move' to the new IRD number. Some extra restrictions are placed on the predecessors: the predecessor IRD number must cease to be used; and a majority of employees must move to the successor in the month immediately following the 'death' of the predecessor IRD number. In addition, certain industries are treated differently: in agriculture, for example, it is common for groups of casual employees to move en masse between farms, so the predecessor firms in agriculture are restricted to a minimum of 20 employees. The aim is to minimize the number of false negative links – where we do not recognize that two employers in different periods are in fact the same- without introducing excessive numbers of false positive links between employers – where two records are linked but in fact refer to different employers. Clerical review showed that the level of false positives was approximately 2 percent, while the level of false negatives was at most 5 percent. Using this method, approximately 10 percent of firm deaths in any month are shown to be continuing businesses; this corresponds to up to 37 percent of job and worker flows due to spurious firm deaths (Kelly, 2003). At the beginning of a tax year (April), the percentage of spurious births is seen to be as high as 17 percent. When comparing these results to those of the probabilistic matching described above it should be noted that the employer repair is carried out only for employers (these are the only firms filing an EMS form) and that employers are a subset of the universe of businesses on the LBF.

Reporting units in the EMS may represent more than one enterprise on the LBF. A change in employer IRD number in this case indicates a change in filer within a group of firms filing jointly, rather than a usual predecessor/successor relationship. Comparison to the links resulting from the probabilistic matching identifies instances of changes in group filing practices.

The US Census Bureau Longitudinal Employer-Household Dynamics Program (LEHD) also identifies continuing employers by tracking groups of common employees between firms (McKinney and Prevost, 2002). The LEHD approach differs from ours in two main ways: we consider only the case where the predecessor IRD number ceases to be used and require the change of employer IRD number to occur in consecutive months. Our alternative source of information for repairing longitudinal links, the probabilistic matching of geographic units on the BF, ensures that few of the links that would be picked up if our criteria for tracking employees were as extensive as LEHD are missed.

4 The LBF

The LBF contains information about businesses from two data sources. Businesses on the LBF may be in both the BF and the EMS data, or in only one. Information about employers on the LBF generally comes from both the EMS payroll data and the BF; except in the case of businesses found only in the

EMS data. There are approximately 17,000 (3 percent) IRD-only businesses on the LBF in any month. IRD-only businesses tend to be small with limited information. These firms have region imputed based on the location of employees and only approximately 50 percent have an industrial classification. Information about non-employers comes from the BF. For units that are represented in both data sources, it is often the case that information about changes in administrative numbers is observed in the payroll data before it is observed in the BF data. This is due to processing lags on the BF. The main exception to this is the movement of geographic units between existing firms. This movement, which is usually due to the sale and purchase of a 'going concern', does not usually involve any observable change in administrative numbers at the legal unit level. Changes such as this are picked up by the BF annually via a frame update questionnaire and appear on the LBF in due course, with the reference month of the change as detailed by the respondent.

The unique identifier on the LBF is the permanent business number (PBN): the PBN remains the same over time even when geographic numbers, enterprise numbers and IRD numbers representing the unit in the data sources change. In a sense, the identifiers currently used in the source data to refer to a business become attributes of the PBN, in the same way as industrial or region classifications are attributes. The smallest unit on the the LBF is the PBN, which represents a geographic unit on the BF or a reporting unit in the payroll data. Enterprises on the LBF may be composed of one or more PBNs, depending on whether the enterprise is a single geographic unit enterprise or a multiple geographic unit enterprise.

Understanding the strengths and weaknesses of the LBF is essential if it is to be used for research.

Strengths

The most obvious advantages of the LBF are that the data are longitudinal and frequent. There are no industry or geographic scope restrictions. Statistics on firm birth and death, the survival and growth of firms and firm size (in terms of persons employed) can be produced by ANZSIC industry, business type, institutional sector, overseas equity and region. Splits, mergers, take-overs and restructuring are all visible in the database.

The LBF population is larger than the population of employers in the LEED database and also larger than the population of firms on the BF. The advantage of, ultimately, having all businesses in New Zealand represented in the database, not just the largest firms, is that it allows researchers to study firm growth from the very beginning of a firm's life cycle. The accuracy of the information about firms, particularly those firms that change structure or legal character, is enhanced by using information about employees, and matching techniques, to confirm the nature of the changes.

It is possible to link the LBF with other administrative data or business surveys. Linking the LBF to other Statistics NZ firm data is relatively straightforward, making it possible to extend the breadth and depth of information for LBF firms. For example, a project is currently underway in Statistics NZ to link annual enterprise data to the LBF to measure firm productivity and performance.

Weaknesses

Data quality is variable. Firms represented in both data sources can be expected to have accurate and timely information. However, the LBF has limited industry and region information for firms that are found only in the payroll data. Alternatively, small firms found in the BF and not in the payroll data may have attributes such as industry codes updated only once every three years, if at all. The database

also has no information about other characteristics of firms such as sales, purchases or expenditures on research and development.

5 Conclusion

Data integration is seen by policy and official statistics agencies as a growth area. It is viewed as a way of creating information-rich datasets with less resource, and with less burden on respondents, than that needed for censuses or surveys. Longitudinally linked firm data provide the potential for analyzing dynamic aspects of economic behaviour. The use of existing administrative data to create the LBF has the advantage that respondent burden is reduced, and the level and complexity of information available to researchers, analysts and policymakers is increased.

The LBF described in this paper is still under development. Additional research into businesses that undergo extended periods of inactivity and fine-tuning of the repair processes may yield further improvements in quality. Future work includes the addition of self-employed individuals. There are approximately 250,000 self-employed individuals not on the LBF. Information on the self-employed comes from annual tax returns so will be less timely than the monthly payroll data.

References

Barnes M and Martin R (2002)
Business data linking: An introduction. *Economic Trends*, 581, 34-41.

Bycroft C (2003)
Employee IRDN Repair. Statistics New Zealand, www.stats.govt.nz.

Carroll N, Hyslop D, Mare D, Timmons, J and Wood J (2002)
The turbulent labour market. Paper prepared for the New Zealand Association of Economists' Conference, Wellington, 25-27 June 2002, http://nzae.org.nz/files/%2316(2)-CARROLL-TEXT.PDF [22 August 2003]

Gill L (2001)
Methods for Automatic Record Matching and Linking and their use in National Statistics. National Statistics Methodology Series No. 25, Office of National Statistics, UK.

Jarmin R and Miranda J (2002)
The Longitudinal Business Database, Centre for Economic Studies technical note, CES-WP-02-17.

Kelly N (2003)
Repairing EMS Employer Longitudinal Links. Statistics New Zealand, www.stats.govt.nz.

McKinney K and Prevost R (2002)
Successor/Predecessor Firms. Technical Report No. TP-2002-04, Statistical Research Report Series, US Bureau of the Census.

Newcombe H and Kennedy J (1962)
Record Linkage. *Communications of the Association for Computing Machinery,* 5, 563-566.

Seyb A (2003)
The Longitudinal Business Frame. Statistics New Zealand, www.stats.govt.nz.

Seyb A (2004)
Repairing LBF Longitudinal Links. Statistics New Zealand, [forthcoming].

Winkler W. (2001)
Record Linkage Software and Methods for Merging Administrative Lists. Technical Report No. RR-2001-03, Statistical Research Report Series, US Bureau of the Census.

About the Author

Allyson Seyb is a Senior Methodologist in the Statistical Methods division at Statistics New Zealand (Statistics NZ). She is currently managing the Data Integration section, which provides methodological support to Statistics NZ projects involving record linkage. Her former work experience is in the area of business survey design. She can be contacted at Statistics New Zealand, Dollan House, 401 Madras Street, Private Bag 4741, Christchurch, New Zealand; tel. +64 3 964 8786; fax +64 3 964 8999; e-mail allyson.seyb@stats.govt.nz.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Architectural Design of a Survey Questionnaire and Respondent Data Repository.
Practical Considerations

Philip Cookson and Jason Sobell

Abstract

This paper will examine the practical requirements for the technical design of a survey questionnaire and respondent data repository capable of efficiently storing, retrieving, and analyzing survey questionnaire and respondent data, and explain the application of the system for facilitating cross-wave and cross-study data analysis of market research survey results.

The design leverages the capabilities of modern database systems such as Microsoft SQL Server 2005, that provide native support for the storage and query of XML documents. Combining the QEDML standard (an XML based document standard for encoding survey questionnaires), with conventional storage of respondent data in relational data tables; it is possible to build a system capable of reporting on survey questionnaire datasets across multiple research studies and data sources using commercially available database reporting tools.

This design will be described in the context of creating a complete (end-to-end) survey automation system, spanning survey questionnaire development, through survey deployment, data collection, aggregation with historic data sets, and analysis/reporting.

Keywords

Market Research; Survey Automation; QEDML; XML; Database design; Questionnaire Repository

1 Introduction

The business case for implementing survey automation systems

Survey automation is a set of processes, tools, and technology that are used to streamline the specification, design, programming, and deployment of survey questionnaires and the data collection, data storage, and analysis of survey (respondent) data sets. The implementation of comprehensive survey automation systems based on open XML standards and leveraging the capabilities of modern database management software has the power to transform the way that survey based market research is conducted. The organizations that can most benefit from the implementation of survey automation systems are those which are engaged in global research, and/or longitudinal (tracking) studies with respondent data span several years.

An effective survey automation system should meet the following key business objectives:

- improve the consistency, quality, efficiency, and cost effectiveness of creating and deploying global market research studies;
- standardize the way in which survey questions are asked, and translated into multiple languages;
- provide a centralized repository of the survey data collected from market research studies in a form that facilitates cross-wave and cross-study analysis and data mining activities; and
- provide an online reporting tool that will allow (experienced) end-users to access and perform basic analysis on the survey data sets.

The business benefits of implementing a survey automation system are:

- Faster end-to-end turn around of survey based market research projects.
- Greater cost efficiencies compared to labor intensive manual processes, through re-usability, repeatability and economies of scale.
- Improved consistency and quality of survey projects through a reduction in sources of (manual) error at key points in the process.
- Enabling of new value added functions such as real-time online reporting of results, tighter project management, analysis of cross-study data sets, and interactive web-based data collection.

The survey automation system described in this paper incorporates a centralized database repository of survey questionnaires stored in the form of XML documents, that is linked to a survey respondent database that is encoded in a standard relational data table. This hybrid approach enables the retrieval of respondent data from multiple waves of a single research study, or from the same (or similar) questions across multiple research studies using the standard capabilities of modern database management software. These data sets may then be exported into a wide variety of specialized data analysis and reporting packages to facilitate cross-wave and cross-study data analysis and reporting.

What is XML and why is it important to the design of Survey Automation systems?

XML (eXtensible Markup Language) is a text markup language for the storage and exchange of structured information, retaining both the content and context of data.

Figure 1 (next page) provides an example of how XML can be used to mark up data that is associated with the publication of newspaper content.

XML is important as a standard for encoding information because it is:

- **Human readable**
  Encoded using ASCII or Unicode (multiple-language support)
- **Future-safe**
  Self-documenting (As long as we have the means to read the storage media)
- **Extensible**
  New tags and language elements may be introduced while maintaining backwards compatibility
- **Ideal as an interchange format between organizations**
  XML documents can be validated against a shared DTD or XML Schema
- **Flexible**
  Standard tools such as XSLT renderers allow multiple destination formats to be automatically generated

Figure 1: Example of XML representation of newspaper content

```
<newspaper>
  <section type='headline'>
    <article id='213'>
      <heading>New England Patriots Kick A**</heading>
      <body>
        <paragraph>
          The New England Patriots have won their third Super
          Bowl in our years with a dominant second half on
          Sunday, wearing down the Philadelphia Eagles 24-21.
        </paragraph>
        <paragraph>
          It wasn't overpowering, and at times it was downright
          ugly.  But it was more than enough to match the Dallas
          Cowboys run    of the 1990s and certify the Patriots
          of coach Bill Belichick and quarterback Tom Brady as
          the NFL's latest dynasty.
        </paragraph>
      </body>
    </article>
  </section>
</newspaper>
```

These attributes make XML the ideal choice for encoding survey questionnaires to facilitate their storage and interchange, as survey questionnaires:

- Consist of a (hierarchical) tree structure
- Implement looping and conditional sections
- Contain meta-data or attributes relating to questions and sections
- Require rendering to many platforms and output formats
- Must be retained for future access if subsequent survey data is to be of value

What is QEDML and how is it used to represent a survey questionnaire?

The proposed solution is based on the QEDML standard. QEDML is an open standard for encoding questionnaire designs with simple, human readable, tags. Based on the XML standard for document markup, QEDML is able to leverage publicly available XML tools to create questionnaires, and to transform these designs into a wide variety of native questionnaire scripting formats for deployment on commercial survey systems, and to automatically generate HTML, PDF or MS WORD formatted questionnaire documents.

QEDML provides complex questionnaire designs portability between different survey programming languages and systems, and is able to generate accurate representations of the questionnaire even with relatively simple survey scripting languages. With QEDML, the dream of "design once" (using re-usable high level questionnaire components), and "deploy anywhere" (Paper, CATI, Web, CAPI) questionnaires can become a reality.

The solution encodes all survey questionnaire and respondent data into an open standard (QEDML) format which is based on the XML standard for information encoding and exchange. The structure of the underlying data is open and clearly documented making it easy for the data to be converted into alternate storage formats, and for other applications to be written to manage the underlying data store.

Figure 2 provides an example of a QEDML snippet showing how XML can be used to mark up a survey questions.

Figure 2: Example of a simplified QEDML snippet encoding of a survey question

```
<Element ElementType="Question" Orientation="Vertical"
List="GENDER" ListColumns="1" Mandatory="false" Name="QID1"
ListOrder="Normal" QuestionType="SinglePunch" >
    <QuestionLabel>
        Q1
    </QuestionLabel>
    <MainText>
        What is your gender?
    </MainText>
    <PostText>
        Choose One:
    </PostText>
</Element>
```

2 Architectural requirements for a Survey Questionnaire Repository

Why create a survey questionnaire repository?

Although the development of a survey questionnaire is a creative, collaborative process that results in a unique questionnaire design that is specific to the business context being researched; the practical reality is that the vast majority of individual questions in any given research study are contained within a larger "library" of survey questions that are commonly used by the research vendor/organization. Indeed, good research practice dictates that there should be consistency in the wording of questions and the use of scales to enable reliable comparisons of respondent data between studies. In the case of longitudinal (tracking) research studies it is common for 85% or more of the survey questionnaire to be identical with the previous wave of the study. This situation creates a compelling case for creating a survey questionnaire repository that can be used as the foundation for creating new survey questionnaires.

The limitations of traditional (tabular) designs for survey questionnaire repositories

One approach to the architectural design of a survey questionnaire repository is to store each question as a data entity (row) within a relational database structure, referenced by a unique Question IDentifier (QID). The key elements of the question, such as question data type (e.g. Numeric), question text (e.g. What is your age in years?), and other relevant meta-data (e.g. a flag indicating that the question is

mandatory); are each stored as a separate field within a question data table. This design provides an efficient way to store and retrieve individual questions.

However, such a simplistic design approach is not a practical solution to the development of a survey questionnaire repository that is to be used as a tool for creating new survey questionnaires. Assembling a questionnaire "one question at a time" is inefficient. The usual approach for creating a new survey questionnaire is to start with a template questionnaire (which is usually the survey questionnaire that was used in the previous wave of the research study, or one that has been assembled from pieces/sections of survey questionnaires used in similar research studies).

To facilitate this type of use it is therefore necessary for the survey questionnaire repository to be able to store archived versions of prior (complete) survey questionnaires, commonly used sequences of questions (sections), in addition to individual questions themselves. These data structures can not easily be accommodated within a traditional relational (tabular) design approach to survey questionnaire repositories.

A new design approach for survey questionnaire repositories leveraging XML

This can be achieved by representing the questionnaire design in the form of an XML document, using a suitable encoding standard such as QEDML. This XML document can then be stored directly in a database as a structured XML document leveraging modern database systems such as Microsoft's SQL Server 2005 that provides native support for XML document storage, retrieval, and access.

During the XML document loading process it is possible to automatically decompose the XML document into its component questions, each of which can be stored individually (as an XML snippet and as a relational data structure).

Because each element of a question is tagged as a separate XML element it is possible to quickly compare these elements (using a simple hash function algorithm, or standard XML comparison utility), with the other questions already stored in the repository, so that "new" questions may be automatically identified and stored in the database.

In this manner it becomes possible to created an archived repository of complete questionnaires, partial questionnaires (i.e. sequences of questions), and individual questions; that can be used as the basis for efficiently and consistently designing new survey questionnaires.

This design approach also has the benefit of being able to store all of the (foreign language) translated text associated with each element of a question, as well as higher level meta-data associated with the project (such as project description, vendor, deployment type, etc.).

The net result is a system capable of quickly loading and retrieving complete survey questionnaire designs and/or individual questions along with all of the relevant meta-data. Note that in practise, such a system will require the development of utility applications that support the editing and manipulation of the underlying data, such as to verify the accuracy of the (foreign) language translations, and merge/delete "almost" identical questions that are loaded into the database.

Implications of using XML for the design of a survey questionnaire repository on the design of the survey respondent data repository

Storing a complete survey questionnaire design in the form of an XML document that can be directly manipulated within the repository database system dramatically simplifies the design requirements for the survey respondent data repository. Survey respondent data sets can be stored as a standard relational data table using a compact data structure, and linked to the survey questionnaire XML document via a Survey Project Identifier (SPID), which provides a unique mapping between a specific (XML) instance of the survey questionnaire. Responses to specific questions in the associated survey respondent data set (for that instance of the survey questionnaire), are linked via a unique Question Identifier (QID) that is common to the survey questionnaire and survey respondent data repositories.

3 Architectural requirements for Survey Data Collection and Reporting

Design Issues for Survey Data Collection versus Survey Data Reporting

The design requirements of survey data collection differ from those of survey data reporting in two key aspects:

(i) Survey Data Collection: The importance of "state" in data collection

For survey data collection it is important to maintain "state" information for each individual respondent to the survey so that their progress through the survey questionnaire can be tracked (and if necessary, re-traced). This is especially important for web-based survey data collection methods, which commonly allow a respondent to pause in the middle of a survey, and then resume the survey at a later date/time (at the same point that they left the survey, with all of their previous answers intact, and with the ability to precisely retrace their steps). The key state information that needs to be stored for each respondent is: *Survey Version ID* (indicating which specific version of the survey a respondent started – since the survey may be modified during the fielding process); *Respondent "Breadcrumb"* (a coded sequence trail of the questions completed by the respondent so far); and a vector of the *Respondent Rotational and Randomization question states* (which specify which rotational and/or randomization order of questions and or list elements that they have been exposed to, so that they can be recalled exactly as they appeared when the respondent resumes their survey). This state information is only of relevance during the data collection process, and is usually automatically managed by the survey data collection system.

(ii) Survey Data Reporting: The impact of data "Augmentation"

For survey data reporting purposes, the original survey questionnaire design is usually insufficient to fully define the respondent data set, since it is common to "augment" the data collected from respondents with additional information about the respondent derived from other sources (especially for surveys conducted using a panel), and for new data variables to be added that are specific to the analysis process (such as adding a weighting variable, or recoding respondent data into a new data variable). Since this augmentation is specific to each survey that is fielded, and normally takes place outside of the survey questionnaire design system, these structural changes to the data set are not reflected in the original survey questionnaire design. It is therefore necessary to implement some form

of a manual process to re-align the survey questionnaire design so that it accurately reflects the final survey data set.

By implementing a survey automation system that represents the survey questionnaire design in the form of a structured XML document, archiving each change to this design as a (separately stored) XML document; and storing the respondent's answers in the form of a traditional relational data table, it is possible to exactly represent and recreate the survey data collection environment, and optimize the survey reporting capabilities.

Ideally the survey data collection process should be driven directly from the XML specification of the survey questionnaire (which is the case in such systems as QEDML Web Server). However, in practise, this will not usually be the case. It will therefore be necessary to manually create a survey respondent data dictionary. This process can be made easier if the original survey design was coded as an XML document (and used to generate the survey data collection script file via an XSLT transform).

Generating output suitable for use with Statistical Analysis software

Using an XML representation of the final (reporting version) survey questionnaire, it is a relatively straightforward exercise to *automatically* generate a data dictionary (syntax) file, that can be combined with a suitably formatted respondent data file to create a meta-data formatted data file that is suitable for analysis using common statistical analysis tools such as SPSS. This is accomplished using a simple (generic) XSLT transform to convert the XML representation of the survey questionnaire to a suitably encoded syntax file. The following figure shows an example of a SPSS syntax file that has been generated from a correspondent XML encoded version of the survey instrument. This syntax file can be "run" against the data set to generate a fully formatted SPSS .sav document that is suitable for conducting advanced statistical analysis. In theory, it is also possible to embed additional meta-data in the XML representation of the survey questionnaire that could be used to generate an SPSS automation batch file to generate a complete set of analysis tables and charts for the survey. In practice, this is only likely to be economically feasible for standardized tracking studies where the analysis and reporting requirements are consistent over time.

4 Conclusions

XML is a natural means of encoding survey questionnaire designs, and has several advantages over more traditional methods of encoding survey questionnaires for survey data collection and reporting purposes.

The most effective means of creating a robust repository for survey questionnaire and respondent data collection and reporting is to use a hybrid design consisting of an XML encoded survey questionnaire (leveraging questionnaire encoding standards such as QEDML), and a standard relational database table structure for the respondent data set. This design maximizes the capability for using the repository both for reporting purposes, and as an archive of the specific context of the data collection (defined by the precise form of the survey questionnaire), since the XML representation of the survey questionnaire is capable of storing all relevant information about the design, including the programming/script code used to define the skip logic for the survey.

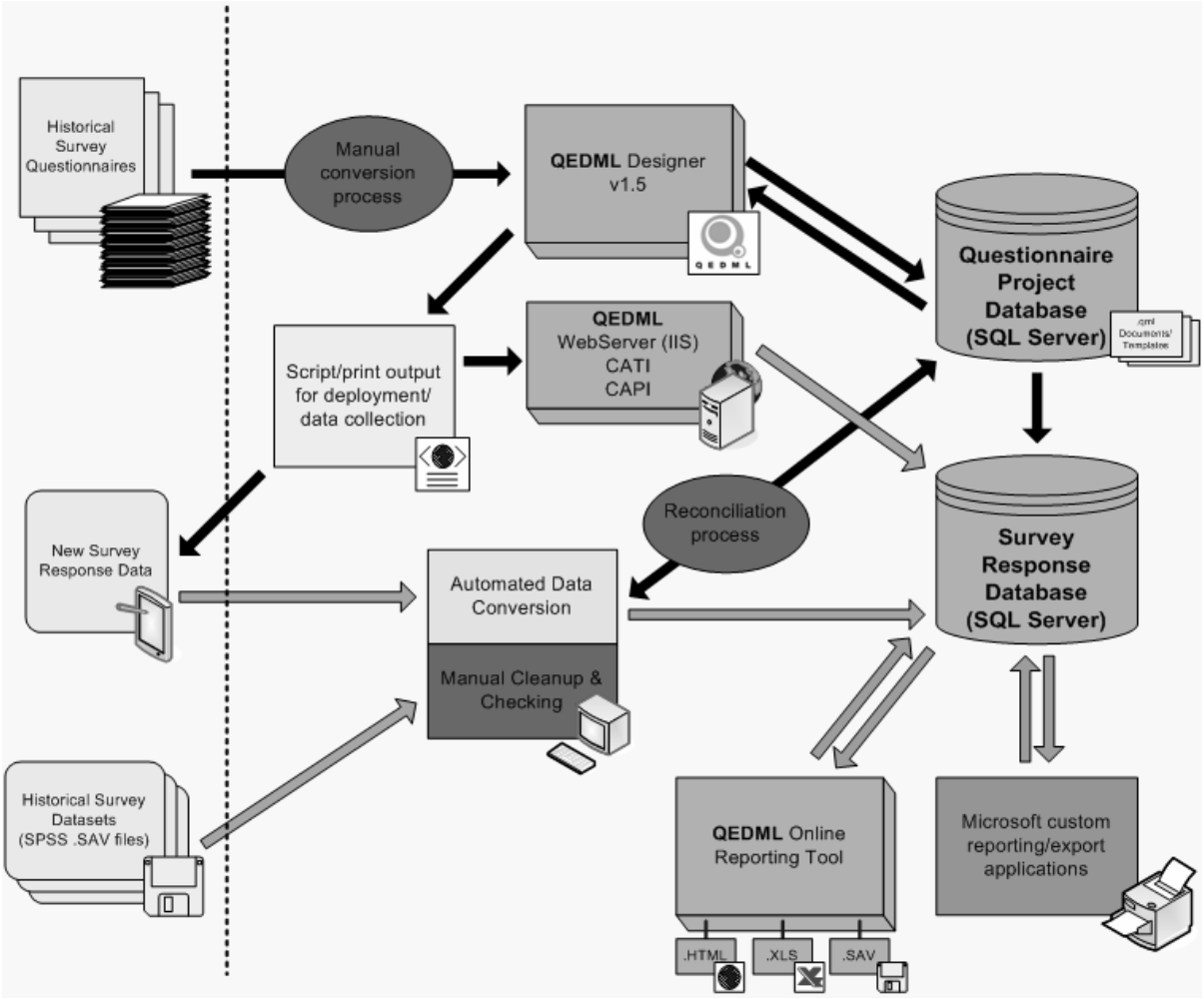
The advent of modern database systems, such as Microsoft SQL Server 2005, that provide native support for XML data structures, enables greater flexibility in the design of next generation survey automation systems based upon this technology.

The design of a comprehensive survey automation system inherently requires the integration of people, processes, and (software) technology. The overall process for conducting "real world" research surveys is not able to be fully automated, since at key points in the process there is always a need to manually augment (i.e. edit, add to, and/or "clean up"), the questionnaire design and the collected respondent data, for example to code open-ended responses or add weighting or analysis variables to the data set. As a result, survey automation software must be supplemented by well documented and coordinated processes and procedures for managing the manual steps involved in the data collection and analysis phases to ensure the data integrity of the overall system.

The architectural approach described in this paper is of particular relevance to organizations that must manage several years of an (evolving) survey research tracking study (where the survey questionnaire changes over time), and organizations that conduct survey research in several different languages (where there is a need to synchronize changes to the survey questionnaire across multiple language variants).

Appendix A: Architectural overview of a real world Survey Automation system

The following diagram provides a very high level architectural overview of a survey questionnaire database and survey respondent data repository system that has been designed specifically to meet the needs of Microsoft's Corporate Market Research function. This design is typical of the survey automation requirements of corporations that conduct a significant number of global market research studies.



Appendix B: Overview of XML support in modern database systems

Modern database systems such as Microsoft's SQL Server 2005 support XML as a native data type, making it possible to create a table with one or more columns of type XML in addition to standard relational data columns. These XML values are stored in an internal format as large binary objects (BLOBs), which enable the XML model to support document order and recursive structures. Because of the seamless integration of the XML data type with SQL Server's storage and programming models, it is possible to query and update XML documents and even write joins between XML data and relational data in the database. The same query engine and optimizer is used for querying XML as for

relational data. With the XML data typed columns, it is also possible to specify an XML Schema Definition (XSD) schema that restricts the XML stored in the column or makes it variable to the vocabulary described in the schema.

At the individual column level, several methods have been added to the underlying SQL language. These new methods enable the searching, retrieving, and updating of data that is embedded within a larger XML structure. Each method may be combined with a XML field to limit or update data associated with an XML column. The five key methods are query(), value(), exist(), nodes(), and modify().

The ***query()*** method allows searching through a larger XML structure to find a set of data based on an XML Query (XQuery) definition. The XQuery language is a World Wide Web Consortium (W3C) standard for searching or defining a set of XML nodes that meet a set of criteria. Most of the other methods also leverage XQuery conditions.

Exist() is an optimized method that lets you screen XML data the same way you screen data with a relational WHERE clause. However, instead of retrieving a value from your XML data, then checking to see whether this value matches a condition, you pass the condition into the XML processor, then retrieve only those records that match the condition.

The ***value()*** method returns a specific value from within your XML structure. The only limitation is that this value must be a single instance, such as a string or number. It can't be a subset (i.e., node) of your XML structure.

The ***nodes()*** method returns a subset of your XML structure in the form of a node. This result can then be used by other methods, such as exist() and value() to pull out repeating values that might be embedded within a single XML column. By using the XQuery syntax, which is native to XML, it is then possible to embed queries against the data in a single relational column.

The ***modify()*** method lets you insert and update values and nodes that are contained within an XML column. The modify() method accepts both INSERT and UPDATE statements not only for scalar values but also for entire subtrees. Thus, it is possible to add specific child elements to a collection of items. By leveraging XQuery statements within each XML column command, it is possible to manipulate the individual components contained within a custom XML structure.

References

There are a range of trade groups and industry publications that focus on XML related issues. Here are few that are most useful for XML in a survey design/development context.

W3C's XML Activity Page (http://www.w3.org/XML/Activity.html)
This is XML central, the place where all new developments relating to the XML standard are recorded.

XML.com: XML From the Inside Out (http://www.XML.com/)
An O'Reilly online publication that focuses on XML.

Microsoft's MSDN XML section
(http://msdn.microsoft.com/library/default.asp?url=/nhp/default.asp?contentid=28000438) A collection of tutorials and other developer relevant information on the XML standard provided by Microsoft.

QEDML Technology Web Site (http://www.qedml.com.au/)
Information and resources on the QEDML (XML) standard for encoding survey questionnaires.

About the Authors

Philip Cookson is an internationally experienced Marketing Researcher, with expertise in the area of Internet-based market research and survey automation technology. Philip is the Director for Server and Tools Product research at Microsoft Corporation, and also serves on the board of directors for Philology Pty. Ltd, a company he help to found. Prior to founding Philology, Philip was Senior Director for Global Market Research & Analysis with Apple Computer Inc., where he was responsible for conducting global market analysis and customer research relating to Apple's new products and strategic initiatives. Philip is a full member of the Market Research Society of Australia, and has achieved QPMR accreditation. Philip holds a Masters Degree in Mathematical Physics, and a Post-graduate diploma in Computer Science.

Jason Sobell is a professionally qualified software engineer and database designer who has extensive international experience as an Information Technology consultant. Jason has a Masters Degree in Information Technology from RMIT University, and lectures in commercial database applications at RMIT University and the University of Melbourne. Jason is co-developer of the QEDML standard for encoding survey questionnaire designs; and is the Chief Technologist at Philology Pty. Ltd. where he is responsible for the technical development of the QEDML product family.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

The Role Of Software As A Value Added Tool in Survey Research

Kevin Wavell

Abstract

This paper will cover some of the issues involved in using and re-using commercial survey data. It will attempt to give some information about a wide range of processes but will not provide an in-depth analysis of the techniques used.

It will explain how the use of information provided by individual respondents can assist in the creation of new survey databases and how the newly-collected results can be linked directly back to the respondents' original data to form an extended dataset.

It will describe how the survey data can form links to other existing data sources via techniques such as data matching, data fusion and attitude modelling, and it will address solutions to the problems arising from international variations in survey design that are critical if comparative global analyses are to be performed.

Finally it will examine ways in which software can facilitate the use of published data by clients.

Keywords

software, BMRB, TGI, Choices, re-contact, database enhancement, international standards

1 An introduction to TGI

TGI is a large, continuous single-source survey with a history going back more than 35 years. It uses a 100+ page self-completion paper questionnaire from a representative sample of 500 face-to-face interviews each week. The survey collects data on all aspects of purchasing, behaviour, attitudes and media consumption and is delivered quarterly on a subscription basis to more than 200 end-users.

The published data is purchased and used by a variety of clients, mainly (but not exclusively) as a media planning tool allowing Media Owners, Advertising Agencies and Brand Owners to gain insight into their marketplace and plan advertising campaigns. Successful selling of TGI means that the potential for further growth of sales in some of these areas is now limited and so TGI is always looking for ways of adding value to the existing data. With 25,000 respondents per annum providing 30,000 items of data, the published survey contains a wealth of information and it makes logical (as well as commercial) sense to encourage maximum use of this information.

The name TGI refers to the Index value that is commonly used to identify key sets of respondents (Target Groups) that are significantly more (or less) likely than average be present in selected cells of

an analysis. Figure 1 shows a simple example of the use of the Target Group Index, and how it might provide a starting point for planning a newspaper campaign aimed at certain Airline travellers.

Figure 1

	elements	total	British Airways: AIR TRAVEL: Airlines Flown Holidays	Easyjet: AIR TRAVEL: Airlines Flown Holidays	Britannia: AIR TRAVEL: Airlines Flown Holidays	Air 2000: AIR TRAVEL: Airlines Flown Holidays
total	Sample	23,479	1,946	1,216	1,094	1,103
	(000)	44,862	4,357	2,784	2,071	2,216
	vert%	100%	100%	100%	100%	100%
	horz%	100%	9.71%	6.21%	4.62%	4.94%
	Index	100	100	100	100	100
Daily Mail						188 365 16.5% 6.49% 131
The Sun	(000)	8,295	534	389	467	242 440 19.8% 5.30% 107
	vert%	18.5%	12.3%	14.0%	22.6%	
The Guardian	Sample	679	121	84	20	28
	(000)	1,242	254	194	36.3	55.8
	vert%	2.77%	5.82%	6.98%	1.75%	2.52%
	horz%	100%	20.4%	15.7%	2.92%	4.49%
	Index	100	210	252	63	91

People who have flown with British Airways are 110% more likely than the average to read The Guardian.

This is expressed as an Index (based on 100) for this Target Group of 210

Index = how much more/less likely?

2 The march of time

The development and growth of TGI has taken place concurrently with the advent of modern computing, so it will be no surprise that the product should have taken full advantage of the hardware and software techniques that have been made available over this time. The transition from punched cards and annual printed reports to scanned questionnaires and quarterly web delivery will no doubt be reflected in many other areas of survey computing.

TGI has also tried to move with the times, not only by keeping its survey content relevant (both to clients and respondents) but also by utilising software to improve and enhance the product itself in a bid to encourage greater use of the data, and thereby maximising its value. The most recent example of this is the response to government initiatives to improve the nation's diet. TGI now calculates and will soon publish a summary of Body Mass Index for each respondent, thus allowing better insight into the habits and attitudes of the key target groups and helping to get the messages to them.

I will not pretend that these developments are wholly altruistic, but the pursuit of increased profitability cannot be achieved without meeting the demands of those who are buying the product. TGI is in a rare, but fortunate position of being able to provide BOTH an industry dataset AND the tools with which to analyse it. As a result we have to continually develop in a bid to add value to the product. So, where have these developments been made?

3 Enhancing and extending the data provided by the respondent

The most obvious way to extend the data that have been collected, making them usable to additional subscribers, is simply to ask further questions in addition to those that have already been asked. The vast amount of data collected on a TGI questionnaire makes it a particularly convenient and cost-effective source for re-contacting minority samples who might otherwise be very difficult to identify.

Eligible respondents can be identified using responses to the main TGI questions and their unique ID numbers are exported to an external file. These ID numbers can then be linked, via a secondary lookup, to a file containing the respondents' personal details (which will also indicate whether the respondent has consented to being contacted again). TGI has been carrying out such work from the early days, when names and addresses were still stored on paper, and the matching was a manual task, so the development of the relational database has certainly had a positive impact on the efficiency of this part of the process. The requirements of the MRS Code of Conduct and the Data Protection Act have to be rigorously followed in the areas of consent and preservation of respondent anonymity, and without the flexible tools provided by modern software, adherence to these principles might be made more difficult. The presence of a respondent database that is completely independent from the questionnaire data allows us to maintain the personal information for all of our contacts, to keep track of the frequency of usage of each individual, and to update consent details if these should change. Importantly, this personal information is only available on a need-to-have basis and is password protected on a standalone PC. Care is taken to ensure that these details can never form part of the same physical dataset as the TGI questionnaire details (or any other survey data).

Recontacting respondents for further research provides valuable data for subscribers, but real added value is gained when these new data are linked together with the responses from the original TGI questionnaire. For the selected sub-sample of respondents, not only are the results of the new research project available for analysis, but so are all of the original responses that were given by each individual. This allows the re-contact questionnaire to focus upon the specific requirements (sometimes too detailed to be asked of a national sample), in the knowledge that the full TGI questionnaire has already been administered and all of these data are also available.

There are times, however, when clients require data sooner than can be made available though a re-contact survey. This lead to the development of the "pre-contact" survey, in which client-specific questions are asked at the initial contact stage. Responses to these questions can then be passed though for linkage with the TGI questionnaire data before publication.

Recontact and pre-contact surveys will typically aim to collect data from areas that are not covered on the TGI itself – these could be detailed questions about a product or brand, more general questions for a smaller regional sample, or measurement of attitudes to more specific issues that are important to a client. The breadth of TGI data means that it is usually unnecessary to ask questions that have already been asked on the TGI questionnaire, so there is little need to deal with potential conflict between answers. However, some surveys have been specifically designed to ask the same questions again with a view to understanding some of the reasons behind changing product purchasing habits or new patterns of media consumption and behaviour. The knowledge of what a respondent did or thought in the past often can provide very good insight into their current decisions and habits when combined with newly collected data.

4 Extending the data by direct linkage to independent respondent data

The techniques described thus far refer to the linkage of data which have a common identifier, and this is most likely to be done where the same fieldwork agency is responsible for collecting both sets of data. But there is also a requirement to link to data that has been collected for the same respondent, but that is available only in an independent external dataset.

BMRB was a pioneer in the field of using survey research data to enhance customer databases. The typical customer database created from purchase history, transaction records or guarantee cards would contain many records (often numbering millions) but would have a very limited set of data, restricted to the specific product purchase or store record history. These records will already have been processed by the client and individuals would be grouped into customer segments using the available data. Although TGI has far fewer records (tens of thousands), each of these records has much broader data coverage. Crucially however, these two very different sets of data will both contain the respondents' name and address - and possibly one or two other basic demographics. Working as a joint venture with dunnhumby, a leading customer relationship marketing company, using standard address matching software a match is carried out, at an individual level, between the TGI respondents and the customer database records to identify records that are present in both datasets. Once this link between the two datasets is established, existing customer segments from the client database can be linked to TGI for modelling and then aggregating back to the whole customer database, enabling it to be profiled using any TGI data. This in effect creates a "virtual" dataset that contains TGI "responses" for many millions of individuals, giving owners much better insight into their customers – what other products and services they buy, their hobbies and interests, their media consumption and their attitudes and motivations. It is an important part of the modelling process that the "real" respondents who formed the original match are removed from the dataset in order to preserve their anonymity. The dataset thus created, now totally modelled, has been shown to significantly improve the effectiveness (and reduce the cost) of targeted direct mailshots.

Many years before customer database enhancement, TGI data was also being extended by the addition of a number of GeoDemographic classifications for each respondent. The principle is simple with the only requirement being the postcode of each respondent. Every postcode in the country has been classified into a typology based upon the characteristics of the neighbourhood making it an easy task to attach the codes to each respondent. Respondent anonymity became an issue here as the GeoDemographic providers requested access to the coded TGI data in order to test and refine their classifications. A simple adjustment to the respondents' ID was all that was required to make a direct linkage impossible and therefore protect the respondent.

5 Extending the data by indirect linkage to independent data

There has been considerable debate over the years concerning the validity of data fusion in our industry. I am not aware of any claim that data fusion can directly predict the answers that an individual respondent would give in the same way that, for example, a direct re-contact of the individual could, but there is sufficient evidence to suggest that the process can and does produce acceptable results at an aggregated level.

BMRB was one of the first companies to explore the potential of commercial data fusion in the early 1990s. After a significant amount of experimentation, and driven by a growing need in the industry, BMRB undertook to create Target Group Ratings by fusing TGI to the BARB Television ratings data.

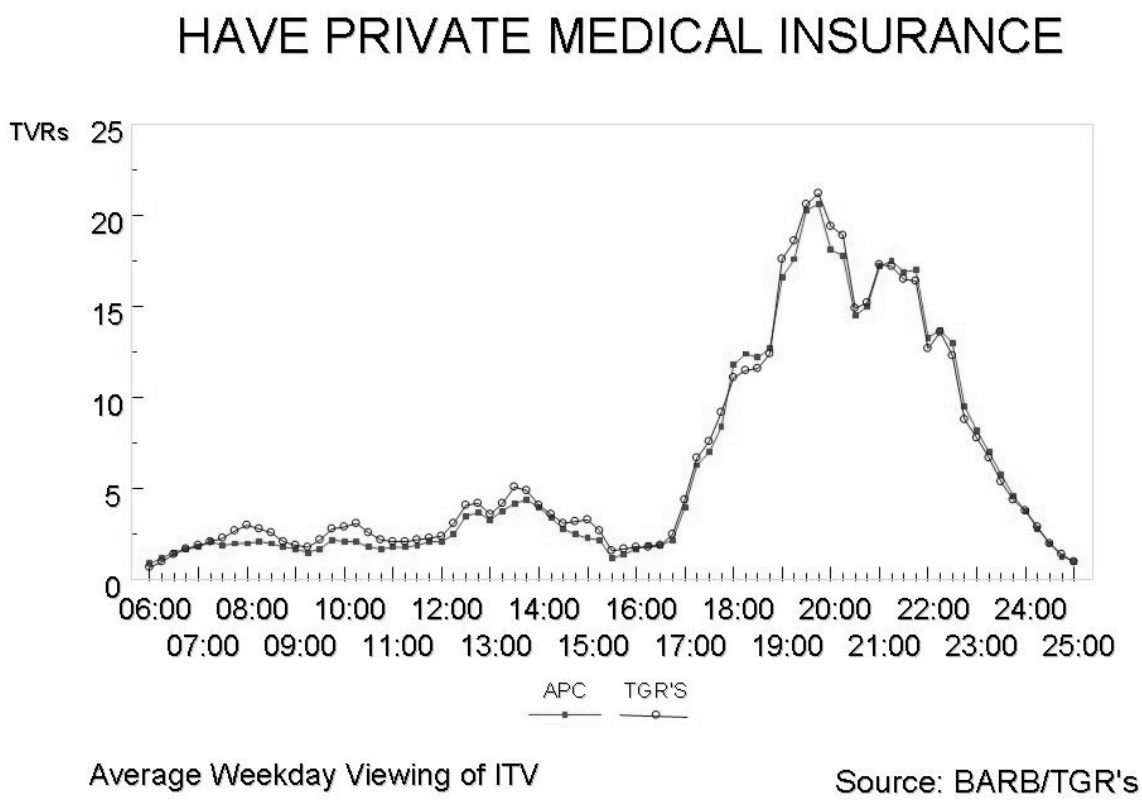
The specific objective was to create a database of TV viewing that would enable TV planners and advertisers to assess the existing BARB ratings by TGI brand and product usage rather than just normal demographics.

An important consideration is any fusion is the selection of the set of common variables that are used to hook the two surveys together. The large amount of data contained in TGI makes it more likely that these variables can be found, and where the relationship between the selected common variables and the donated data is strong, the fusion is usually found to be good.

Independent validation of the output of the initial TGI/BARB fusion by Ken Baker was commissioned by BARB to reassure them of the quality of the output. A repeated independent check of the algorithm was later undertaken on behalf of BMRB by fusing TGI data "to itself" and analysing both the real and fused data. Both set of results were extremely reassuring and Baker has stated that "the TGI/BARB fusion is almost as accurate as it would have been if single source data had been collected".

Further to this, it is actually possible to compare a small number of fused items from TGI with similar original respondent data from the BARB panel classification. There are a limited number of cases where data exist on both sides and will allow such comparisons can be made but figure 2 shows an example of TV viewing behaviour for individuals with Private Medical Insurance (defined from the 2 sources). The same patterns exist elsewhere.

Figure 2



Where possible, BMRB continues to fuse external surveys with TGI, thus extending the use of its data and providing actionable data for clients. Without a genuine fused dataset, evidence suggests that users are forced into "intuitive fusions" by creating surrogate target groups. When users need to analyse the

habits of Bottled Lager Drinkers, they might instead use ABC1 Men aged 18-34. The fusion process provides a more relevant and accurate technique.

The techniques used for linking independent datasets by matching respondents on a common set of variables can also be used to deal with "missing data" within a dataset. This ascription technique is used to impute responses to sections of a questionnaire for which no answers exist, using data that do exist. This technique is well established in many countries and has been used on the US Census since the 1960s.

Respondents for BMRB's Internet Monitor are recruited on the weekly Omnibus survey. This means that around half of all Internet Monitor respondents have also completed the TGI survey which is recruited from the same source. There is, therefore, a significant core of questions common to both surveys which allows the use of ascription techniques to produce data for internet behaviour and attitudes of all TGI respondents who are also internet users. The two completed sets of data are merged to produce a separate product - TGI.Net - a database that allows you to analyse internet behaviour and attitudes by product and media consumption from TGI.

6 Indirect linkages and bridges to external datasets

So far, I have dealt with some of the methods used to facilitate the reuse of data provided by TGI respondents by extending the data in a number of ways at the level of the respondent.

It is a routine requirement for almost every survey to reflect as closely as possible the population from which it was sampled (or which it claims to represent). As a result most survey research findings will include some element of weighting, using the simple technique of applying a numeric value to each respondent in order to correct for imbalance or under-representation. This technique has the effect of allowing the survey data to be treated as if it were the population, and therefore allows easier comparisons to be made.

It is essential for the Print Media Industry in Great Britain to have a common set of data that both buyers and sellers of advertising can use in negotiations (i.e. a currency). This has meant that for many years it has been a requirement for reported TGI readership data for each print title to show the same published figures as the National Readership Survey (where titles are measured on both surveys). This has led to the development of an extremely sophisticated weighting algorithm designed to match the demographic profiles (age and social class – within gender) of more than 200 titles to the equivalent NRS figures. Once the weighting has been applied, TGI and NRS effectively become the same survey at this level, and the populations represented on the NRS can now be examined in detail by using the equivalent TGI data. The two sets of data, while still retaining their independence, can both provide detailed information about readers of titles. NRS can provide the very detailed readership data not present on TGI, whilst TGI has the product and brand data not available on NRS.

Another bridge between independent datasets has been developed after recent work between BMRB and Millward Brown. This project has used a set of carefully selected attitude agreement statements to segment the TGI data based upon respondents' broad priorities in life. This is based upon well-established psychological evidence (Schwartz's model of universal values), and has 2 underlying dimensions (Conscience and Spirituality vs. Self Interest and Image, and Safety and Conservatism vs. Adventure and Exploration). Based upon each individual's position on these dimensions relative to the national average, each individual is assigned to one of nine clearly distinctive groups. Each of these groups has demonstrably different values and characteristics. By adding these same attitude statements

to Millward Brown's continuous brand tracker (ATP), and using exactly the same algorithm to segment that data, a bridge is created between the two datasets. The analyses that can result from the Life Value Segments on both sets of data will give a much deeper understanding of the consumers' motivations and values, and their media and leisure habits on TGI will indicate opportunities for communication and sponsorship. Figure 3 shows ATP reporting awareness of the marketing of Car C to be greatest among the Independents and Experiencers groups; figure 4 then examines TGI data for sports viewing on TV within these two groups to identify potential opportunities for sponsorship.

Figure 3

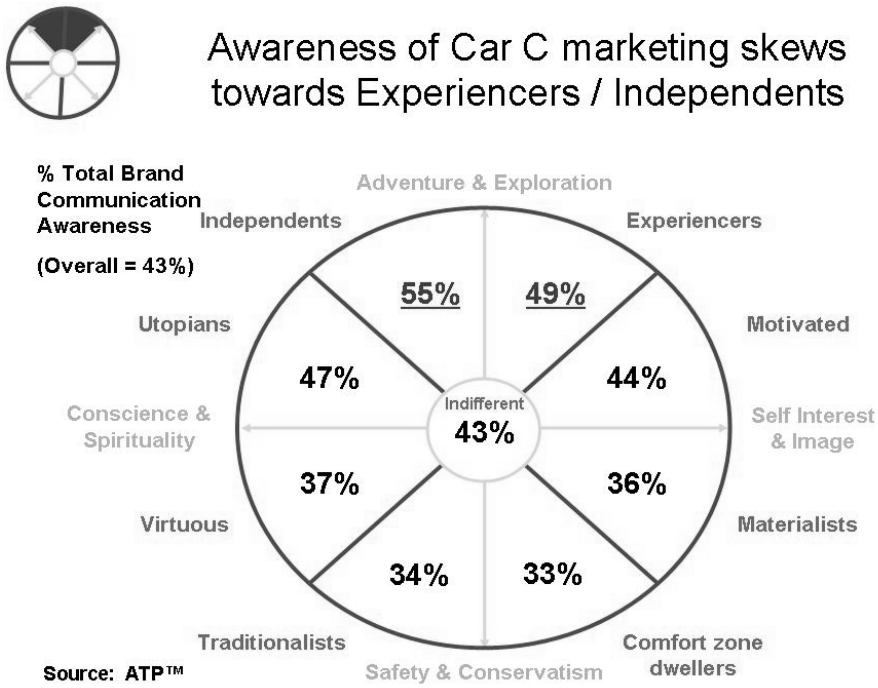


Figure 4

Linking to TGI - Sponsorship opportunities can be explored for example...

Sports like to watch on TV:	Total car target	Experiencers Index vs total car target	Independents Index vs total car target
Football	41.9%	109	110
Snooker	29.7%	80	98
Rugby Union	29.5%	105	117
Athletics	28.7%	112	92
Cricket	28.5%	92	112
Tennis	27.6%	106	84
Motor Racing	27.4%	120	101
Golf	23.7%	108	98
Rugby League	21.7%	105	122

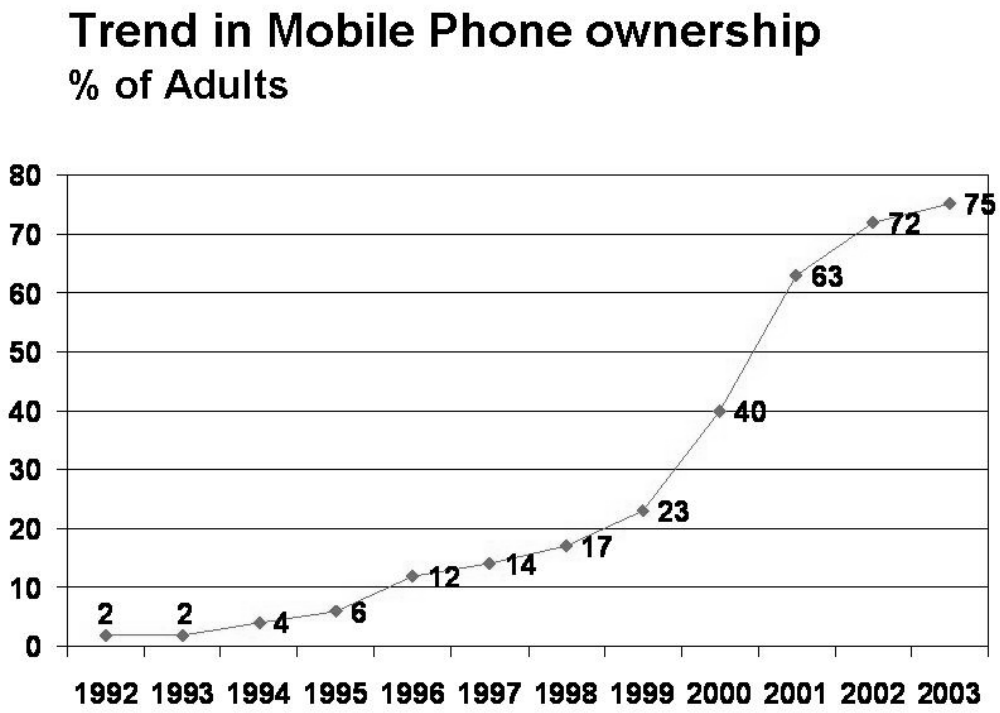
Source: TGI

7 The presence or absence of metadata

The development of TGI's own PC-based software, Choices, in the 1980s, created an analysis system where the data were independent of any fixed location. This was achieved by using a relatively simple metadata structure that relied upon the allocation of a unique 8 character keyword to every data item. The metadata structure was organised hierarchically to allow easy navigation of the questionnaire and all data were accessed using an internal indexing system that was linked to the unique keyword.

By keeping the metadata relatively simple, we were able to create similar metadata for other surveys and allow these to be delivered into Choices, both as standalone surveys or, where relevant, linked to a TGI study (as mentioned earlier). The key to the success of the system (and the TGI data) was that the unique keyword, once allocated, remained permanently linked to its data item. Although the metadata (and the software) have become enriched over time, the keyword linkage remains intact and the philosophy is unchanged since the first release of the software. By allowing Choices to access multiple survey databases, it becomes an easy task for users to perform trend analyses with no knowledge required of any aspect of the data itself. Simply selecting a set of data from the tree menu in one survey, and requesting the analyses to be carried out over selected historical releases was sufficient to produce a trend analysis. Figure 5 shows the value of being able to easily access historic TGI data in a single run.

Figure 5



Source: GB TGI 1992 -2004

The fixed common keyword approach worked extremely well for all surveys that we processed and published under our control, and was subsequently also used to produce comparative analyses across databases for Great Britain, Northern Ireland and the Republic of Ireland (which all shared the same core keywords when the latter surveys were launched in the early 1990s). But it soon became obvious

that we could not use a similar technique for comparisons between other geographical datasets as TGI rapidly expanded internationally (now in 58 countries). The process was made more complex as a result of the differing levels of harmonisation between the surveys, because a number of partner countries already possessed surveys that were similar enough to TGI to make it sensible to co-operate with (rather than set up in opposition to) these research agencies. Although data from all of the TGI Global network partners are capable of being loaded into and delivered using Choices, in some cases alternative proprietary metadata exist which need to be converted into Choices format before the process can be completed. A number of the alternative metadata systems are not based around a keyword system, so no time based analyses are possible.

The role of Global TGI is to encourage greater use of the data by including it as part of a multinational dataset, and the lack of access to historic data is not such a major concern. Globally, it is more important that we are able to identify and report common items of data (particularly brand usage data) across a range of geographical databases that may have no initial common metadata and no form of data linkage (keyword or otherwise). The solution was to develop a tool that was capable of recognising the one single identifier that we could totally rely upon being present in every database that we needed to analyse – a descriptive text. This Multi Survey Search tool will identify all relevant text matches (plus near matches and associated words) and then output full details relating to the context of the matches into a spreadsheet. The spreadsheet details can then be customised and saved in the form of an analysis specification.

The process works extremely well and Global brands such as Coca Cola and McDonalds can easily be found with no prior knowledge required of the dictionary or database structure. So, even without metadata, text fragments and associations have been able to provide a good link between disparate international datasets adding greatly to their use and reuse.

8 How software can encourage users to get more from the data

By far the greatest efforts in software development over the last decade have been focussed on helping users to get the best from the data that they have. It is no longer sufficient to just provide simple crosstabulation tools and whilst we should not ignore the need for software to be intuitive and easy to use, particularly for infrequent users, it is increasingly the tools that can inform and help in the decision making process that are in greatest demand. TGI has certainly provided its share of these tools in line with our belief that TGI should be providing insight and not just data.

In the late 1980s TGI launched Correspondence and Cluster Analysis on a desktop PC. Because of the high machine specification required (in 1989 a Compaq 386 with 110MB HDD and 9MB RAM cost more than £7,500), analyses were initially restricted to internal use, but demand was such that it became a part of the distributable software within a few years. With these tools, supplemented by CHAID, users now create their own market segmentations.

It was soon increasingly obvious that the market craved segmentations of the population that were not restricted to the standard combinations of demographics. TGI began to deliver standardised market segmentations on a regular basis, from the Financial Marketplace to the identification of Early Adopters. TGI has also been keen to explore opportunities with other developers who might have alternative ideas about adding value to the data that we deliver. By the late 1990's Choices had linked to the "Grande Mappa" (Big Map) technique developed by Eurisko. Two recent developments with Pointlogic have led to the creation of new products based around the TGI dataset. Although very

different in their application, both products allow an element of subjectivity to be included into the mix of variables selected for analysis. Targetware makes the creation of target groups more realistic by allowing the user to adjust the relative importance of each component variable via a sliding scale. Compose, a tool primarily aimed at channel planners, allows the combination of data from a number of sources, and again uses a screen based approach to the selection and relative importance of criteria.

9 Helping users to understand the data

Like all businesses, TGI has adapted to the advances in desktop software. A team of TGI Client Service and marketing executives are continuously digesting the data that we collect, in a search for actionable information for clients (as well as selling the benefits to potential subscribers). Regular Centre for Excellence courses are held which teach subscribers how the data can and should be used in order to maximise their investment. Articles regularly appear in the press and the TGI website regularly updated with stories that have been told and published using TGI data – and, yes, this service is freely available to anybody to cares to look at www.bmrb-tgi.co.uk.

As mentioned earlier when talking about Body Mass Index, we certainly do not see TGI as a static product, and we attempt to add value to each release. We have made international comparative analyses much easier by the creation of a common Global Socio-Economic Level segmentation, based upon a defined basket of data items that are asked on all of our Global questionnaires. We have created segments of Consumer Expenditure Groups and Liquid Asset Scores based not solely upon Income, but taking into account savings and expenditure.

All of these segmentations are created from source data and added to the published dataset for general analysis.

Time marches on relentlessly, and TGI will undoubtedly continue to utilise software improvements and develop specialist user tools in an attempt to broaden the use and awareness of the product. It is critically important to TGI that the data is not just sold, but that it is used creatively and on a regular basis, so the package that we sell includes data, software, training and support. TGI will continue to innovate and provide users with the insight and information needed to drive their businesses for many years to come.

About the Author

Kevin Wavell is Technical Director of BMRB/TGI Surveys. He has been involved in the area of survey data processing for more than 30 years, initially with Pritchard Brown and Taylor before moving to Pulse Train and then to Demotab in 1978. He joined BMRB in 1986 and has been closely involved in the processing, delivery, development and growth of TGI at all levels since then. He now advises and trains TGI publishers both domestically and internationally on all aspects of data collection, processing and publishing and is closely involved with TGI's software development teams.

Methodology and Software for Complex Models

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Modelling Complexity in Health and Social Sciences. Bayesian Graphical Models as a Tool for Combining Multiple Sources of Information

Nicky Best, Chris Jackson, Sylvia Richardson

Abstract

Researchers in substantive fields such as social, behavioural and health sciences face some common problems when attempting to construct and estimate realistic models for phenomena of interest. The available data tend to be observational rather than collected via carefully controlled experimentation, and are typically fraught with missing values, unmeasured confounders, selection biases and so on. These features often render the use of standard analyses misleading; instead a comprehensive set of inter-dependent submodels are needed to model the data complexities and core processes that researchers want to understand. It is also invariably the case that a single dataset fails to provide all the necessary information, and many complex research questions require the combination of datasets from multiple sources. Bayesian graphical models provide a natural framework for combining a series of local submodels, informed by different data sources, into a coherent global analysis.

This paper introduces the key ideas behind Bayesian inference and graphical models in this context and shows how they can be used to easily construct models of almost arbitrary complexity. The ideas are illustrated by two case studies involving the integration of survey data, census data and routinely collected health data. Analysis of graphical models such as those presented here can be carried out using the WinBUGS software for Bayesian modelling.

Keywords

Conditional independence; data synthesis; epidemiology; hierarchical models.

1 Introduction

Applied statistics is about making sense of empirical observations and maximising the information content that can be extracted from data. The modern discipline is being presented with increasingly challenging problems as technological advances allow the collection and storage of vast quantities of data – ranging from the micro level of the human genome to the macro level of, for example, geographically-indexed health data, urban transport networks or climate data – and researchers and others wish to use such data to answer ever more complex questions. It is also invariably the case that a single dataset fails to provide all the necessary information, and many complex research questions require the combination of datasets from multiple sources. Faced with these challenges, the applied statistician needs a set of conceptual and computational tools to enable him/her to capture the essential

structure of a complex problem and to maximise the amount of useful information about this that can be extracted from the data to hand. In this paper, we aim to show that the techniques of *graphical modelling* offer a natural and coherent framework both for building complex statistical models that link together multiple data sources, and for drawing inferences from them in order to deal with real world problems.

A key idea underpinning the specification of a graphical model is that of conditional independence, and in Section 2 we explain this link in more detail. In Section 3, we discuss how graphical models and conditional independence assumptions provide a natural way of building complex statistical models from a series of simple local submodels. This is illustrated by the first of two case studies, which shows how multiple sources of data can be linked together to address a problem in environmental epidemiology. Section 4 provides a short overview of the Bayesian approach to statistical inference and the simulation based algorithms used to carry out Bayesian computation. The central role of graphical models in facilitating these computations will be emphasised, and the WinBUGS software – which makes use of all these concepts – will be introduced. In Section 5, we present a second case study that uses the graphical modelling approach to investigate the link between socioeconomic factors and ill health, and makes use of both individual level and aggregate (small area) level sources of data. We end with a discussion in Section 6.

2 Conditional independence and graphical models

Graphical models consist of nodes representing the random quantities in the model, linked by directed or undirected edges representing the dependence relationships between variables. A simple example involving four random variables W, X, Y and Z, connected by directed edges, is shown in Figure 1(a). (next page) Such models have many forms, including path analysis diagrams which are used extensively in structural equation modelling (Dunn et al, 1993), Bayesian networks and their use in probabilistic expert systems (Lauritzen and Spiegelhalter, 1988), and more recently, causal diagrams that provide conditions for making causal inference from empirical observations (Pearl, 1995). In all of these cases, graphical models are used to provide a pictorial representation of the relationships between random variables, and in particular to encode conditional independence assumptions underlying a statistical model. At one level therefore, graphical models provide a qualitative visual description of the model structure without the need for complex algebraic formulae. Such pictures provide a valuable tool for communicating the essentials of a complex model to a wide audience. In addition, however, the conditional independence assumptions represented by these graphs provide a formal mathematical basis for deriving a joint probability distribution for the random variables in the model, which leads directly to a statistical model.

To make these ideas concrete, consider again the graphical model in Figure 1a. The arrows in the model imply that both W and X depend directly on Y and Z, but the absence of a link between W and X implies that, conditional on Y and Z, W and X are independent. This means that once the values of Y and Z are known, discovering X tells you nothing more about W. Now suppose that the variables W, X, Y and Z represent genetic information – say blood group – on different individuals. By the standard laws of Mendelian inheritance, one's blood group directly depends (probabilistically) only on one's parents' blood groups. Hence Y and Z could represent the blood groups of two parents, and W and X could represent the blood groups of two of their children. If the parents' blood groups are known, then knowing child X's blood group provides no additional information about his/her sibling

W's blood group. Of course, if the parents' blood groups are not completely known (i.e. *un*conditional on Y and Z), then X's blood group *is* informative about W's blood group.

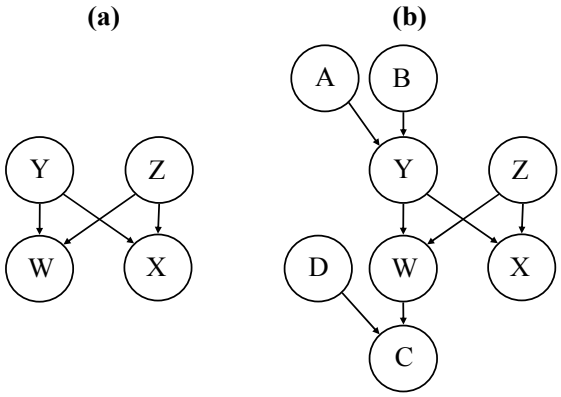


Figure 1. (a) Simple graphical model showing conditional independence relationships between four variables; (b) Elaborated graphical model showing relationships between eight variables.

Figure 1(b) shows a more complex graph for eight variables. Continuing with the genetic example, this graph could represent the conditional independence relationships between the blood groups of four generations of a family, with nodes A and B representing the blood group of two grandparents, and node C representing the blood group of their great-grandchild, who has parents W and D. In general, the nodes in a graphical model can represent any observed or unobserved random variables of interest in a particular problem, and not just genetic quantities. In this case, the directed links between nodes will often represent known or supposed causal relationships. For example, in a graphical model of an epidemiological study of lung cancer, we might include a directed link from a node representing the variable 'smoking' to a node representing the variable 'lung cancer'. Undirected links are also possible. These represent association or correlation between variables, rather than 'cause-effect' relationships. However, for simplicity, we will focus only on directed graphs in this paper.

Whatever the nodes in a directed graphical model represent, it is convenient to use the general terminology of 'parents', 'children' etc. when considering the formal mathematical properties of these graphs. In particular, it can be shown that the joint probability distribution of all the quantities (nodes) in the graph has a simple factorisation

$$p(V) = \prod_{v \in V} p(v \,|\, \mathsf{parents}[v])$$

where v denotes an arbitrary node, V denotes the collection of all such nodes in the graph and the notation p(A|B) denotes the conditional probability distribution of variable A given the value of variable B. This is an extremely powerful result which says that we only need to consider the relationship between each node or variable in our model and its parents (direct influences) – and in particular, specify the conditional distribution of each node given its parents – in order to fully specify the joint distribution and hence the statistical model. The task of writing down a complex joint probability model for a particular problem is thus simplified into one of specifying a series of 'local' 'parent-child' relationships between each variable and its direct influences. For example, the joint distribution (i.e. probability of any particular combination of blood groups for the eight individuals) represented by the graph in Figure 1(b) is

$$p(A, B, C, D, W, X, Y, Z) = p(A)\, p(B)\, p(Y|A, B)\, p(Z)\, p(W|Y, Z)\, p(X|Y, Z)\, p(C|W, D)\, p(D)$$

The next section provides a more detailed illustration.

3 Building complex models

In Section 2, we introduced two key ideas – conditional independence and the factorisation theorem associated with directed graphical models. Here we present the first of our two case studies to show how these ideas can be used to help build complex statistical or probability models by splitting up a large system into a series of smaller components, each of which contains only a few variables and is easily comprehensible.

Case study 1

This case study is based on an epidemiological study currently being undertaken in the authors' department to investigate the risk of low birthweight associated with mothers' exposure to water disinfection byproducts. Chlorine is routinely added to the municipal water supply in the UK as the main means of disinfection. This serves an important public health purpose; however, the added chlorine also reacts with naturally occurring organic matter to form a range of unwanted byproducts, the most widely occurring of which is a group of compounds known as the trihalomethanes, or THMs. Some studies have found that exposure to high levels of THMs is associated with increased risk of adverse birth outcomes, such as low birthweight, still birth or congenital defects, although the evidence is inconclusive (Nieuwenhuijsen et al, 2000). Since only a small percentage of babies are born with low weight (< 2.5kg), large sample sizes are needed to investigate any such relationship. Our study therefore uses routinely collected data for the whole of Great Britain, with cases and denominators obtained from the national births register. Data on THM concentrations have been obtained from routine tap water samples taken by 14 water supply companies in Great Britain for regulatory purposes, and can be linked to births via geographic identifiers (postcode grid reference). Whilst the large sample size provides us with high statistical power and the data are cheap and routinely available, these data have a number of limitations. The THM measurements are sparse, with some areas and time periods having no observations; they also relate to THM concentrations in the tap water, whereas a pregnant mother's personal exposure to THM depends on factors such as how much tap water she drinks, whether she filters the water first or drinks bottled water, and how often and long she bathes or showers for (THMs can be absorbed through the skin or via inhalation as well as by being ingested). Various estimates relating to these activities and the associated uptake of THMs have been published in the literature, and we would like to make use of these in this study. The routine births register also contains very limited information on other potential risk factors and confounders for low birthweight, such as ethnicity or smoking. A second data source on maternal factors and birth outcomes is also available to us – the Millenium Cohort Study (MCS). This contains detailed information on around 20,000 babies born in the year 2000, including their birthweight, plus a rich source of information on parental factors that may be confounders in our study. However, the MCS lacks sufficient power on its own to address the question of whether exposure to THMs increases risk of low birthweight. The series of graphs in Figure 2 show how we can construct 'local' submodels for different aspects of this study, using different data sources, and then link these together into a complex global model. The key idea being exploited here is that conditional on certain variables (which may be

observed data and/or unobserved data or parameters), one set of variables in independent of another, and so a modular approach can be taken to building the global model.

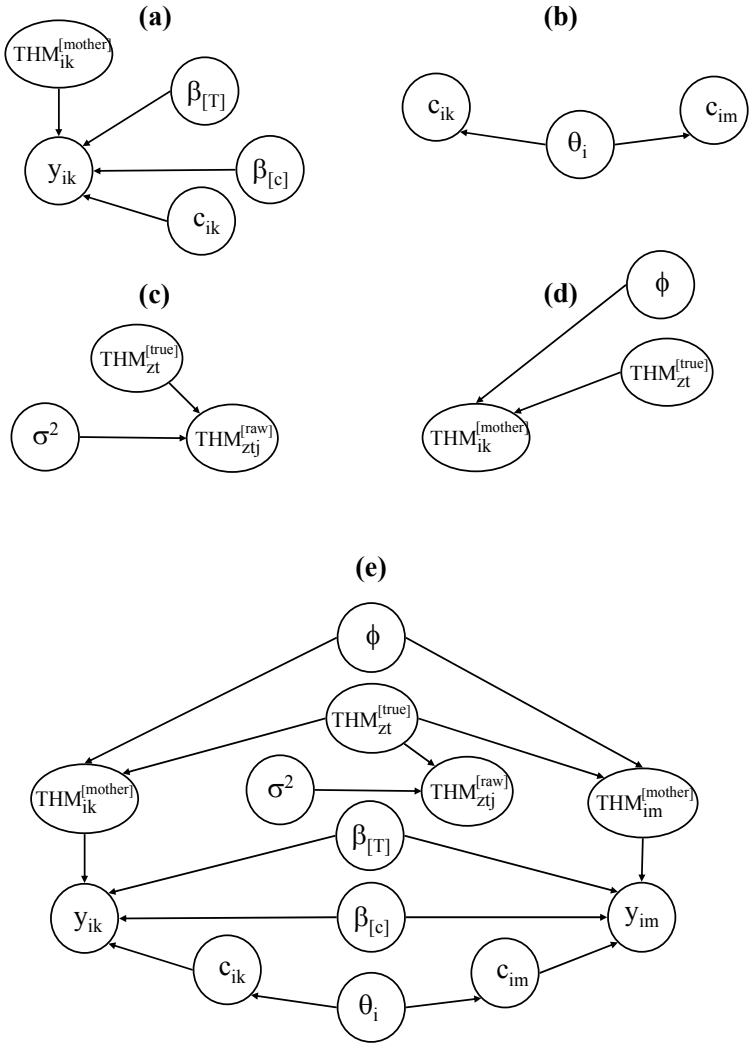


Figure 2. Graphs used to build the model for Case Study 1: (a) epidemiological submodel relating birth outcome to exposures and confounders; (b) missing data submodel describing distribution of unmeasured confounders in national data; (c) measurement error submodel relating measured and true tap water THM concentrations; (d) personal exposure submodel relating true tap water THM concentration to mothers' uptake of THMs; (e) full model created by linking submodels together (note that sub-model (a) is used twice here – once for the national data using subscript k to index individuals in each area i, and once for the MCS data using subscript m to index individuals in area i).

Figure 2(a) shows the *epidemiological submodel* relating occurrence of low birthweight, denoted by the binary indicator, y_{ik}, for baby k in group i (we return to the definition of 'group' later), to the mother's uptake of THMs during pregnancy, $THM^{[mother]}_{ik}$, other risk factors/confounders, c_{ik}, and their associated regression coefficients $\beta_{[T]}$ and $\beta_{[c]}$ (where the latter represent the log odds ratios of low birthweight associated with each risk factor compared to the baseline). A standard statistical analysis would typically specify a logistic regression model to relate the birth outcome to the covariates. If we

assume that the distribution of y_{ik} conditional on its parents in Figure 2(a) is Bernoulli, with logit-transformed rate equal to a linear combination of $THM^{[mother]}_{ik}$, c_{ik}, $\beta_{[T]}$ and $\beta_{[c]}$, this would represent exactly the same logistic regression model. The idea of the graphical model representation is to separate the essential structure of the model from the algebraic detail (although the latter still needs to be specified in order to make inference from the graphical model).

Note that the same epidemiological submodel can be specified for both the national data and the MCS data (for clarity, we will use the index k for babies and mothers in the national data and index m for babies and mothers in the MCS). However, whereas the MCS data contain full information on key confounders of interest – that is, c_{im} is fully observed – this information is missing for the national data. We therefore specify a second submodel, the *missing data submodel*, to estimate the missing confounders c_{ik} in the national data. Figure 2(b) shows one possible graph for this submodel. Here we assume that the distribution of the confounders of interest (e.g. smoking, ethnicity) can be stratified according to group characteristics (indexed by i) that are measured in both the MCS and national datasets – for example, the area of residence. This is plausible, since both smoking rates and ethnic mix are known to show strong geographical variations. The quantities θ_i in Figure 2(b) can be interpreted as the group level proportions or mean values of the confounders in group i, which directly influence the individual values of confounders in both the MCS and national data. The conditional independence assumptions represented by this graph also provide a mechanism by which the observed values of c_{im} in the MCS can be used to make inference about the missing values of c_{ik} in the national data. Thinking back to the genetics example in Section 2, knowledge of a child's blood group provides some information about his/her parents' blood groups when these are not known. Hence the c_{im} provide information about θ_i, which in turn provides information about c_{ik} (although if we knew θ_i, say from another data source, then c_{im} would add no further information about c_{ik} according to the conditional independence assumptions expressed in the graph in Figure 2(b)).

The graph shown in Figure 2(c) represents our third submodel, the *measurement error submodel* relating $THM^{[raw]}_{ztj}$, the measured THM concentration in the j^{th} tap water sample for water supply zone z and time period t to the true average tap concentration for that zone and period, $THM^{[true]}_{zt}$. This graph represents the structure of a classical measurement error model, whereby the observed measurement depends on the true value with some error that has variance denoted σ^2 in Figure 2(c). Again, when the true values are unknown, the raw data, or 'children' in the graph, will provide information about them. In fact, the actual measurement error model we have developed for the THM data is somewhat more complex than that represented in Figure 2(c), involving mixtures of distributions for different water sources, and assuming that the true values in each zone and period themselves depend on further unknown parameters representing the true average concentration across all zones supplied by a particular water source (Whitaker et al 2004). The graphical model is easily elaborated to represent these features.

Figure 2(d) shows the *personal exposure submodel*. This model relates a mother's personal uptake of THMs during pregnancy, $THM^{[mother]}_{ik}$, to the true average THM concentration in her tap water during that period, $THM^{[true]}_{zt}$, and parameters ϕ representing the distribution of personal factors (such as time spent showering, amount of bottled water consumed) likely to affect an individual's exposure to THMs. The latter are not known for each mother in either the national or MCS datasets, and so must be randomly sampled from plausible distributions based on published results from the literature.

Finally, we can combine the four submodels to give a single global model as shown in Figure 2(e). Notice that the linking is done by identifying variables or nodes that appear in more than one submodel, and that conditional on these nodes, the variables in one submodel are independent of the

variables in another submodel. It is this property that allows us to build a complex global model in a simple modular way which links together multiple data sources. Having done this however, we still need an inferential framework and computational algorithms to allow us to learn about the quantities of interest in the model on the basis of the empirical data we have observed.

4 Bayesian inference and computational algorithms

Various approaches are possible for making statistical inference from graphical models. When all the nodes in the graph represent *observed* random quantities it is usual to estimate the parameters of the probability distributions underlying the graph using classical methods such as maximum likelihood. However, when the parameters of these distributions, as well as observed quantities (the data) and unobserved but potentially observable quantities (such as missing data, mis-measured data) are all explicitly represented as nodes in the graph – as in the case study in Section 3 – a Bayesian approach becomes the natural inferential procedure. This is because Bayesian inference is based on assuming that both the observed data and all unknown quantities in the model are random variables with associated probability distributions. In contrast, classical methods of inference treat the parameters as fixed but unknown, and only the data are assumed to be random variables with associated probability distributions. The Bayesian perspective of assigning probability distributions to unobservable quantities such as model parameters has led to much controversy. Here we simply emphasise that by treating parameters and other unknown quantities as random variables, the Bayesian approach allows probability distributions to represent *uncertainty* about the true value of these quantities. It does not mean that parameters have to be viewed as repeatable or variable quantities, or that they have to represent potentially observable events. Viewed as a tool for using probability statements to quantify uncertainty about quantities of interest, the Bayesian paradigm offers a very powerful and flexible approach to inference.

Bayesian inference is based on a straightforward manipulation of conditional probability. As before, let V denote the set of all variables in our graphical model for which we have specified a joint distribution $p(V)$ as the product of parent-child relationships. If we now split V into two parts, with Y denoting all the variables that have been observed in our dataset(s), and θ denoting the remaining unobserved quantities, then our inferential goal is to calculate the conditional probability distribution of θ given Y. According to Bayes theorem, this is given by

$$p(\theta \mid Y) \propto p(Y \mid \theta) p(\theta)$$

where '$\propto$' denotes 'is proportional to', $p(\theta)$ is termed the prior distribution and reflects our uncertainty about the unknown quantities prior to including the data, $p(Y \mid \theta)$ is the likelihood which specifies how the observed data depend on θ, and $p(\theta \mid Y)$ is the posterior distribution which represents our uncertainty about the unknown quantities θ *after* taking account of the data. Note that the right hand side of the equation above is just a standard factorisation of the joint distribution $p(V) = p(Y, \theta) = p(Y \mid \theta)\, p(\theta)$, so if we can write down $p(V)$ we can write down the form of the posterior distribution (up to proportionality). This posterior distribution forms the basis for all our inference. However, being able to write down the equation representing the posterior distribution is not sufficient, and we will usually want to summarise the distribution in some way (for example, to obtain point and interval estimates for specific elements of θ from it). Such summaries involve integrating $p(\theta \mid Y)$, which is potentially very difficult or impossible to do analytically. Instead, simulation-based techniques such as Markov chain Monte Carlo (MCMC) algorithms have been developed to carry out complex integrations. Such

methods work by generating a large sample of values of θ from the posterior distribution of interest, and then calculating appropriate numerical summaries of these samples to approximate the required summaries of the posterior distribution. For example, the mean of the sampled values of an element of θ is used to approximate the posterior expected value of that variable.

There are numerous issues to do with MCMC algorithms that are beyond the scope of the present paper to discuss. The interested reader is referred to Gilks et al (1996) and Brooks (1998) for accessible introductions to the field. The one aspect that we wish to touch on briefly is the link between graphical models and a particular MCMC method known as Gibbs Sampling. This algorithm generates samples from the joint posterior distribution $p(\theta|Y)$ by generating values for one element of θ at a time from the conditional posterior distribution of that element given fixed values of all the other elements of θ. These conditional posterior distributions can be derived directly from the graphical model, using another property of directed graphs which states that the distribution of a node v conditional on all other nodes in the graph just depends on the nodes which are parents or children of v or other parents of v's children. In complex models with thousands of nodes, this leads to considerable simplification of the distributions sampled from by the Gibbs Sampler, and also provides an automatic rule for constructing these distributions if the model is specified as the product of parent-child relationships implied by the graphical representation. The WinBUGS software is a general-purpose Bayesian modelling package that implements Gibbs Sampling. WinBUGS directly exploits the properties of graphical models discussed above, in terms of both how the model is specified by the user, and how the conditional distributions needed by the Gibbs Sampler are internally constructed by the software. The program includes a graphical interface that allows the user to specify their model by drawing the corresponding graphical model. Alternatively, the model can be specified in text format by expressing each of the parent-child relationships corresponding to the graph in the BUGS language. WinBUGS currently has around 15,000 registered users worldwide and is freely available from `www.mrc-bsu.cam.ac.uk/bugs`.

5 Case study 2

In this section, we present a second case study to illustrate the model building strategy and Bayesian inferential approach discussed above. This study is concerned with addressing questions about health inequalities and the socioeconomic determinants of disease. There is a large body of evidence pointing to geographic differences in rates of major illnesses such as heart disease and cancer in the UK and elsewhere, and often these differences reflect geographic patterns of socioeconomic deprivation. In attempting to understand these trends, one important question is the extent to which the socioeconomic gradient of ill-health depends on individual-level risk factors of the people living in deprived areas or on contextual effects or characteristics of the areas themselves. For illustration, consider a simple scenario where we wish to study the effects on risk of developing heart disease (denoted by the binary indicator y) of an individual risk factor such as smoking (denoted x) and an area-level indicator of the level of deprivation in the neighbourhood where each individual lives (denoted Z). One study design to investigate this question is to use individual-level data on health outcomes and individual risk factors, and build a multilevel model with individual and area level effects. The graph in Figure 3(a) shows such a model for the scenario described above, where k indexes individuals living within areas indexed by i. We introduce some additional notation in this graph. Square nodes indicate quantities that are regarded as constants rather than random variables (and so are not assigned probability distributions but are simply conditioned on when specifying the joint distribution represented by the

graph). The large rectangles labelled 'person k' and 'area i' denote repeated structures called 'plates' – that is, all nodes enclosed within a particular plate are repeated for all units indexed by the plate label; nodes outside a plate are not repeated, but if they are directly linked to nodes within a plate then the links (arrows) will be repeated for every plate. The nodes in the graph in Figure 3(a) have the following interpretation. $\beta_{[0]}$, $\beta_{[x]}$ and $\beta_{[Z]}$ are regression coefficients associated with the baseline risk, smoking (x) and deprivation (Z) respectively, and α_i denotes an area-specific residual that captures the differences in risk of heart disease between areas that are not explained by individual smoking habits or the area deprivation – that is unmeasured contextual effects associated with risk of heart disease. These α_i parameters are often termed random effects or random intercepts in the multilevel modelling literature. Finally, σ^2 represents the between-area variance in these residual risks. Note that Figure 3(a) implies that, conditional on the variance, the area-specific residual risks are independent of each other. If we suspect that areas close together may have more similar risks of heart disease than areas further apart (for example, due to shared unmeasured risk factors that have not been explicitly included in the model), then this assumption is not reasonable, and the graph and underlying model would need to be extended to allow for spatial dependence between the α_i's in different areas.

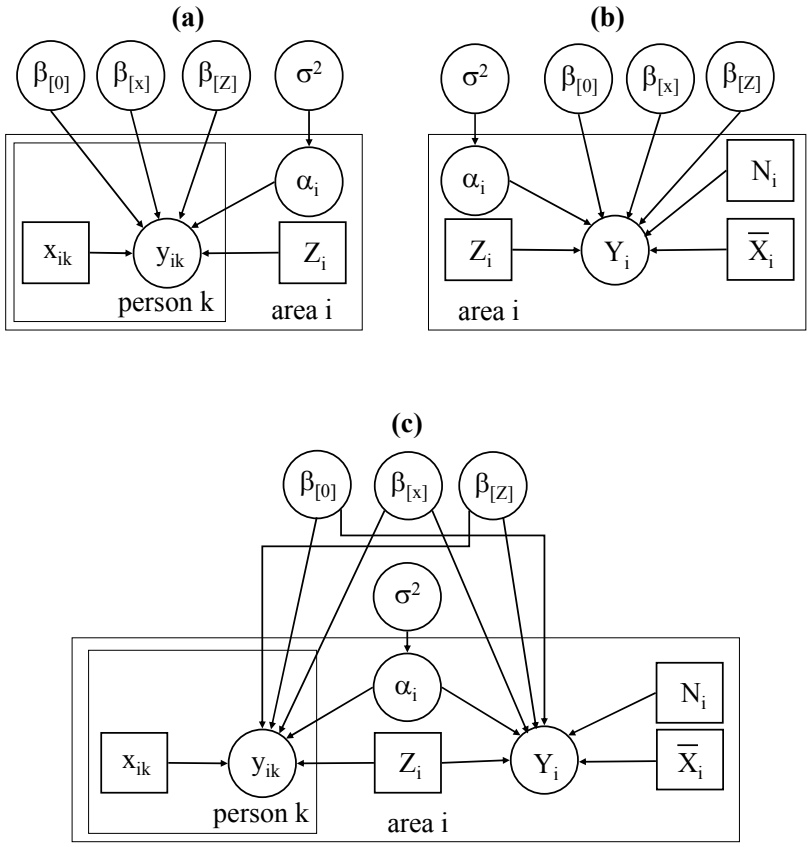


Figure 3. Graphs used to build model for Case Study 2: (a) multilevel model for individual data; (b) ecological model for aggregate data; (c) combined model for mixed individual and aggregate data.

In order to make inference from this model, we need to specify probability distributions for each of the parent-child links represented in the graph. The y_{ik} are binary indicators of disease, which suggests the following distribution for y_{ik} given its parents

$$y_{ik} \sim \text{Bernoulli}(p_{ik}) \text{ where } \text{logit}(p_{ik}) = \beta_{[0]} + \beta_{[x]} x_{ik} + \beta_{[Z]} Z_i + \alpha_i.$$

Here, $\beta_{[0]}$ represents the average log odds of disease in the baseline group in the study region, α_i is an area-specific residual representing the log odds ratio of disease in the baseline group in area i compared with the whole study region, and $\beta_{[x]}$ and $\beta_{[Z]}$ are the log odds ratios of disease associated with the corresponding risk factors compared with the baseline group. A convenient choice for the distribution of the random effects given their parents is

$$\alpha_i \sim \text{Normal}(0, \sigma^2).$$

To complete the model specification we need to specify (marginal) prior distributions for nodes at the 'top' of the graph that do not have parents, that is $\beta_{[0]}$, $\beta_{[x]}$, $\beta_{[Z]}$ and σ^2. We do not give details here, but these distributions can either be chosen to be vague, or if suitable prior information is available, this can be utilised to specify an informative distribution for some or all of these parameters.

One difficulty with the study design just described is that the data available for estimating the area or contextual effects lack power. Most survey or cohort datasets containing relevant individual-level information on health outcomes and risk factors will have only a few (or often no) individuals living in any particular small area. Therefore, an alternative study design is to use routine data sources such as the census and disease registers, which provide information on socioeconomic risk factors and health outcomes for the whole population, but are only available at an aggregated level – for example counts of disease cases per small area, or the proportion of individuals claiming housing benefit in an area. Ecological regression models can then be used to relate the area level average values of risk factors to the rate of disease in each area. The graph in Figure 3(b) represents such an ecological model for the heart disease scenario considered above. Here, Y_i and N_i are the number of cases of heart disease and the population living in area i respectively, $\overline{X}_i$ is the proportion of smokers living in area i, and Z_i is the area deprivation score as before. The remaining quantities in the graph have the same interpretation as in the model in Figure 3(a).

Ecological regression models have been criticised because they can suffer from many types of bias (e.g. Greenland and Robins 1994). In particular, the group level association between the exposure and outcome of interest is not necessarily the same as the individual level association between the same variables – something known as the ecological fallacy or ecological bias. This can be at least partially addressed by fitting a more complex regression model to the aggregate data, which involves integrating the corresponding individual level model within each area (e.g. Wakefield and Salway 2001). Hence in order for the regression coefficients $\beta_{[0]}$, $\beta_{[x]}$, $\beta_{[Z]}$ to have the same interpretation in both the individual and ecological models, we need to specify the following distributions for the parent-child relationships represented in Figure 3(b):

$$Y_i \sim \text{Binomial}(N_i, q_i) \text{ where } q_i = \int p_{ik}(x_{ik}, Z_i, \alpha_i) f_i(x) dx$$

with $\alpha_i \sim \text{Normal}(0, \sigma^2)$ and prior distributions on $\beta_{[0]}$, $\beta_{[x]}$, $\beta_{[Z]}$ and σ^2 as before. Here $p_{ik}(x_{ik}, Z_i, \alpha_i)$ denotes the probability or risk of disease for an individual k with covariate value x_{ik} who lives in area i, and $f_i(x)$ denotes the distribution of covariate x within area i. The details of this integration are unimportant, other than to point out that we need to specify the algebraic details of the model appropriately. However, even with an appropriately specified model, unless there are big contrasts between areas in the values of the risk factors (i.e. unless $\overline{X}_i$ varies considerably across areas), aggregate data will contain little information for estimating the regression coefficients of interest. Our goal, therefore, is to use a mixed study design that combines both the individual-level and aggregate-

level data sources, with a view to improving inference about individual and contextual effects over what can be learned from one source alone.

Figure 3(c) shows how the multilevel model for individual data in Figure 3(a) and the ecological model in Figure 3(b) can be combined to give a single global model. Jackson et al (2005) describe these models in detail, and present a comprehensive simulation study to demonstrate the advantages of the combining data sources using the linked model compared to the individual or ecological models alone.

Application to the analysis of hospital admissions for heart disease

As a brief illustration of the type of inference that can be drawn from these models we consider a small example looking at the effect of three individual-level socioeconomic characteristics (household access to a car, social class, and ethnicity) and area level deprivation on risk of being hospitalised for heart disease. Aggregate counts of hospital admissions for 759 electoral wards in London were obtained from the Hospital Episode Statistics database for 1998, and demographic and socioeconomic covariate information for the same wards was obtained from the 1991 UK census. Individual-level data on both the health outcome and covariates were obtained from the 1998 Health Survey for England for a sample of 4463 individuals living in London. In addition, the Sample of Anonymised Records, which is a 2% sample of individual records from the 1991 census, referenced by the district of residence, was used to provide additional information on the joint distribution of the three covariates of interest within wards (all wards in a given district were assumed to have the same joint distribution). This information is needed to carry out the integrations necessary for the ecological model. Table 1 shows the results of fitting models to (i) the individual-level data only; (ii) the aggregate data only; (iii) the individual-level and aggregate data combined. These three models correspond to those shown in Figures 3(a)-(c) respectively, but without including the area-level random effect. The final column in Table 1 shows the results of fitting the full combined model shown in Figure 3(c) including an area-level random effect to capture any residual contextual effects. Models were fitted using the WinBUGS software.

Table 1. Estimated odds ratios (95% uncertainty intervals) for the effects of socioeconomic variables on risk of hospitalisation for heart disease, estimated using different models and data sources.

	Individual data	Aggregate data	Combined data	Combined data + random effects
Area deprivation	1.00 (0.95, 1.06)	0.99 (0.98, 1.00)	0.99 (0.98, 1.00)	0.99 (0.98, 1.00)
No car access	0.93 (0.55, 1.56)	0.78 (0.71, 0.85)	0.79 (0.72, 0.86)	0.80 (0.61, 1.00)
Low social class	1.27 (0.73, 2.23)	1.07 (0.85, 1.34)	1.12 (0.92, 1.38)	1.20 (0.69, 1.80)
Non white	3.96 (2.38, 6.59)	4.36 (3.96, 4.81)	4.33 (3.94, 4.77)	3.70 (2.70, 5.00)

We note the wide 95% uncertainty intervals for the estimated odds ratios from the individual-level data alone. The aggregate data alone provide tighter uncertainty intervals which partially overlap with those from the individual data, although there are clearly discrepancies between the point estimates from the two data sources. This may partly reflect bias in the aggregate data and lack of power in the individual data. In simulation studies (Jackson et al, 2005), we have shown that the combined analysis tends to yield estimated odds ratios that are both less prone to bias and have smaller mean squared error than those based on a single data source. Finally, when area-level random effects are included, the uncertainty intervals for the odds ratios increase. This is to be expected, since the random effects model accounts for clustering or dependence of the response data within areas. We can quantify the amount of variation in risk of hospitalisation for heart disease that is due to contextual variation (i.e. variation between areas) compared to individual-level variation within areas by calculating the variance partition coefficient (VPC; Goldstein et al, 2002). This is a function of the random effects variance, σ^2, and the sampling variance of the data, and has an interpretation similar to the intra-class correlation coefficient. This shows that only around 5% of the variance is due to unexplained area-level factors, suggesting a relatively small contribution of contextual factors to risk of hospitalisation for heart disease.

6 Discussion

In this paper, we have aimed to show how graphical models can provide the building blocks for linking together multiple data sources in a flexible and coherent way to allow complex yet realistic models to be developed and analysed. We have emphasised the close connections between the graphical representation of the model structure, the fact that this lends itself naturally to a Bayesian interpretation of the model, and the computational methods for making inference using simulation-based MCMC algorithms. We have also noted that the WinBUGS software provides readily available tools to facilitate the specification and estimation of Bayesian graphical models. Many of these ideas are also discussed in a paper by Spiegelhalter (1998).

Even if one did not want to adopt a fully Bayesian approach to data analysis, the ideas discussed in this paper can still provide useful tools for thinking about complex models. At a more informal level, graphical models can be used simply to help represent and communicate the structure of a model, and to guide the model building process by breaking down a complex global model into a series of simpler submodels. One may then chose to estimate each submodel separately using standard statistical methods where available, and conditioning on the estimated values of the nodes that link one submodel to the next. Indeed, this is the approach we have so far adopted for Case Study 1, where the sheer size of the dataset (some 2-3 million births in total) prohibits a single global analysis using MCMC methods. Instead, we have estimated each submodel separately – for example, the measurement error model has been used to estimate the true THM concentrations in the tap water for each mother. These estimates have then been plugged in to the personal exposure submodel (treating them as if they were known values) to generate predicted personal uptakes of THMs for each mother, which are then plugged in to the epidemiological submodel, and so on. The difficulty with this multi-stage approach to inference is that uncertainty about the estimated parameter values from one submodel is ignored if point estimates are then plugged in to the next submodel, and this will yield overly-confident estimates of final quantities of interest. This can be partially addressed by conducting a sensitivity analysis to different values of the plug-in estimates for each submodel. By contrast, if the

full joint model is estimated simultaneously, as in Case Study 2, uncertainty about all the unknown quantities in the model will be correctly propagated.

References

Brooks S (1998)
Markov chain Monte Carlo method and its application. *The Statistician*, vol. 47, 69-100.

Dunn G, Everitt B and Pickles A (1993)
Modelling Covariances and Latent Variables using EQS, Chapman & Hall, London.

Gilks W, Richardson S and Spiegelhalter D (1996)
Markov chain Monte Carlo in Practice, Chapman & Hall, London.

Goldstein H, Browne W and Rasbash J (2002)
Partitioning variation in multilevel models. *Understanding Statistics*, vol. 1, 223-231.

Greenland S and Robins J (1994)
Ecological studies – biases, misconceptions and counterexamples. *American Journal of Epidemiology*, vol. 139, 747-760.

Jackson C, Best N and Richardson S (2005)
Improving ecological inference using individual-level data. Submitted. Available from www.bias-project.org.uk.

Lauritzen S and Spiegelhalter D (1988)
Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, vol. 50, 157-224.

Nieuwenhuijsen M, Toledano M, Eaton N, Fawell J and Elliott P (2000)
Chlorination disinfection by-products in water and their association with adverse reproductive outcomes: a review. *Occupational and Environmental Medicine*, vol. 57, 73-85.

Pearl J (1995)
Causal diagrams for empirical research. *Biometrika*, vol. 82, 669-710.

Spiegelhalter DJ (1998)
Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics*, vol. 47, 115-133.

Wakefield J and Salway R (2001) A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A*, vol. 164, 119-137.

Whitaker H, Best N, Nieuwenhuijsen, M, Wakefield J, Fawell J and Elliott P (2005)
Modelling exposure to disinfection by-products in drinking water for an epidemiological study of adverse birth outcomes. *Journal of Exposure Analysis and Environmental Epidemiology*, vol. 15, 138-146.

About the Authors

Nicky Best is Reader in Statistics and Epidemiology at Imperial College, London. She carries out methodological and applied research in social and health sciences, with a particular focus on small area methods and on Bayesian approaches to modelling complex sources of variability in medicine and epidemiology. She is part of the team developing the WinBUGS software for Bayesian analysis, and is Director of the Imperial College 'BIAS' node of the ESRC National Centre for Research Methods (NCRM) which aims to develop Bayesian methods for combining multiple individual and aggregate data sources in observational studies (www.bias-project.org.uk). She can be contacted at:

Dept of Epidemiology & Public Health, Imperial College Faculty of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG; tel 020 7594 3320; fax 020 7402 2150; email n.best@imperial.ac.uk.

Sylvia Richardson is Professor of Biostatistics at Imperial College, London, and has worked extensively on Bayesian methods and spatial statistics applied to medicine, epidemiology and genetics. She is a collaborator with Nicky Best on the BIAS node of the ESRC NCRM.

Chris Jackson is a Research Associate at Imperial College, London and is currently working on the BIAS node of the ESRC NCRM.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

MCMC Estimation for Random Effect Modelling. The MLwiN Experience

William J. Browne

Abstract

Multilevel models and their extensions to other random effect models that account for the underlying dependence structure of the data when modelling have become very popular in many application areas. There are several methodological approaches to fitting such models and two such approaches that are available in the MLwiN software package. These are likelihood based methods using the iterative gemeralised least squares (IGLS) algorithm and Bayesian methods using Monte Carlo Markov chain (MCMC) methods. In this paper we will give some background on these methods and describe previous work on comparison of their performance. We will then discuss the ease of model extension offered by MCMC methods. We will comment on how we have utilized this ease of extension through an incremental approach that has been used in developing the MCMC functionality in MLwiN. With this in mind we will describe two particular extensions to the multilevel model, cross-classifications and multiple membership models and contrast the ease of expansion when using MCMC with the difficulties in extending the IGLS algorithm. We will finish by discussing further work that is currently starting in various research projects being run by several academics associated with MLwiN. This work includes multilevel factor modelling, models with responses at various levels, missing data through multiple imputation and sample size calculations for complex random effect models.

Keywords

Multilevel modelling, Monte Carlo Markov chain, random effects

1 Introduction

Over the past 25 years there has been an increasing interest in fitting realistically complex statistical models to the large datasets that are often found in the social and medical sciences and other application areas e.g. surveys in official statistics. Such models account for the underlying complex structure by including random effects to represent the structure of the dataset. One of the first application areas to embrace this approach was education. Here the underlying structures were typically nested (or hierarchical) for example pupils nested within classes nested within schools and several multilevel software packages such as MLwiN (Rasbash et al. 2004) were developed to fit such structures. A typical two level model with pupils nested within schools can be written as follows:

$$y_{ij} = \beta_0 + X_{ij}\beta_1 + u_j + e_{ij} \qquad (1)$$

$$u_j \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2)$$

Here *i* indexes pupils and *j* indexes schools. In this model we have a regression relationship between our response *y* (e.g. exam score) and our one predictor *X* (e.g. intake score) and we have random effects u_j for each school that move the population regression line by an amount u_j for school *j*. The model has two variances, a between schools variance σ_u^2 and a residual variance σ_e^2, and is often called a variance components model. These two variances mean that we cannot use standard matrix formulae (as with a regression model) to estimate the unknown parameters. To fit these models we therefore need to use either an iterative procedure e.g. IGLS (Goldstein 1986) or simulation based methods e.g. Gibbs sampling (Gelfand and Smith 1990). Such two level structures appear in many other application areas e.g. health (patients nested in hospital) and surveys (people nested in local areas).

The structure of the remainder of this paper is as follows: In section 2 we describe briefly how both these approaches can be used to fit models such as model 1 above. We then discuss previous work on simulation studies that compare the performance of the methods on such models and also on multilevel logistic regression models. In section 3 we then contrast how these two approaches can be extended to fit further models for more complex data structures, in particular cross-classified and multiple membership models. We show how the MCMC methods are far easier to extend and describe a notation and diagrams that are useful for this class of models. In section 4 we then discuss several other extensions that may be of interest to researchers working on large and complex datasets and are currently being developed by researchers associated with the MLwiN software package.

2 Multilevel Models

The term multilevel model does not have a fixed definition and generally encompasses some subset of possible random effect models. Here we are considering nested random effect models with any number of levels and any number of random terms at each level of the model, for example for a general 3-level normal response model we could write

$$y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk}v_k + Z^{(2)}_{ijk}u_{jk} + e_{ijk} \qquad (2)$$
$$v_k \sim MVN(0,\Omega_v), u_{jk} \sim MVN(0,\Omega_u), e_{ijk} \sim N(0,\sigma_e^2)$$

with the 3 levels indexed by *i,j* and *k*. Here *y* is the response variable, *X*, $Z^{(3)}$ and $Z^{(2)}$ are predictor variables with random effects at levels 1,2 and 3 represented by *e,u* and *v* respectively. In this section we look at two alternative algorithms for fitting such models. As can be seen extending the above notation to many more levels becomes cumbersome and we will discuss alternatives in section 3.

IGLS and RIGLS Estimation

The IGLS algorithm is an iterative maximum likelihood based method that can be used to fit multivariate normal response models with structured covariance matrices of the general form $Y \sim MVN(X\beta, V)$. Here for a dataset of *N* observations we have an *N*N* variance matrix *V* which is described by a few parameters that produces a structured matrix. For example in the variance components model (1) we can find an equivalent multivariate normal model where we describe *V* in terms of the two variance parameters σ_u^2 and σ_e^2. Here $V_{ii} = \sigma_u^2 + \sigma_e^2, V_{ij} = \sigma_u^2$ if *i* and *j* belong to the same school and $V_{ij} = 0$ otherwise. Then assuming the data is sorted by schools *V* will be a block diagonal matrix.

The IGLS algorithm consists of two steps, firstly updating β conditional on the current matrix V by generalized least squares (GLS) and then updating the parameters in V conditional on the current values of β again by GLS having equated expectations of the cross-products of residuals to the V matrix (See Goldstein 1986 for full details). It should be noted that the equivalence of this multivariate response model to model 1 requires the additional constraint that $\sigma_u^2 \geq 0$ which can be incorporated into the IGLS algorithm but will lead sometimes to a boundary solution where $\sigma_u^2 = 0$. The algorithm is fast as it uses fast inversion routines designed to invert the block diagonal variance matrix V. It gives maximum likelihood estimates that are known to be biased for small samples and so an alternative Restricted IGLS (RIGLS) algorithm is available (see Goldstein 1989) that gives restricted maximum likelihood estimates that remove these biases.

MCMC Estimation

We can extend the variance components model (1) to a Bayesian random effects model by adding in prior distributions for each unknown parameter ($\beta, \sigma_u^2 \sigma_e^2$) as follows

$$
\begin{aligned}
&y_{ij} = \beta_0 + X_{ij}\beta_1 + u_j + e_{ij} \\
&u_j \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2) \qquad (3) \\
&\beta \sim p(\beta), \sigma_u^2 \sim p(\sigma_u^2), \sigma_e^2 \sim p(\sigma_e^2)
\end{aligned}
$$

If these prior distributions are chosen to be 'diffuse' then we should get an approximately equivalent model to model 1. Alternatively we could also incorporate any actual prior information we have to produce informative prior distributions for our unknown parameters. To fit such a Bayesian model we can use a group of methods known as MCMC methods, and in this case a method known as Gibbs sampling. Ideally we would like to make inferences from the joint posterior distribution, $p(\beta, u, \sigma_u^2, \sigma_e^2 \mid y)$ of all unknown parameters given the data but it isn't feasible to do this directly for this model so instead we find the full conditional distributions of the unknown parameters. The Gibbs sampling algorithm then involves sampling in turn from each full conditional distribution, conditioning on the current estimates of the other parameters. This procedure will then converge to sampling (dependently) from the joint posterior distribution.

For model (3) above we will have four sets of full conditionals which given starting values for each parameter can be sampled from in turn. These are $p(\beta \mid y, u, \sigma_e^2), p(u \mid y, \beta, \sigma_u^2, \sigma_e^2), p(\sigma_u^2 \mid u)$, and $p(\sigma_e^2 \mid y, \beta, u)$ and if we choose conjugate prior distributions these conditionals will have standard distributional forms. MCMC methods are computationally more intensive than the IGLS and RIGLS methods but give as output the full posterior distribution rather than simply point estimates and standard errors. Ideally we would like to use MCMC methods when we have no prior information and there has been much research on choosing suitable 'diffuse' or 'non-informative' prior distributions for such situations. We will discuss our input to this debate in the next subsection.

Comparison of Methods

Browne and Draper (2000, 2004) performed large simulation studies to investigate both bias and interval coverage properties (i.e. how often confidence intervals contained the true value) of the IGLS, RIGLS and MCMC estimation methods. For MCMC various choices of 'diffuse' prior distribution

were used using both MLwiN and the WinBUGS package (Spiegelhalter et al. 2000). The simulations centred around a 2-level education dataset with pupils nested within 48 schools. They considered two models, a variance components model similar to model 1 but excluding the effect of a predictor (X_{ij}) and an extension of model 1 to a random slopes model where the effect of the predictor is allowed to vary across schools. These two models were chosen so that 'diffuse' priors for both single variances and variance matrices could be considered.

The basic dataset structure had 864 pupils nested within 48 schools and this structure was studied along with subsets with 6, 12 and 24 schools, balanced numbers of pupils per school and unbalanced numbers that reflect the actual data. The main findings of this research were that perhaps unsurprisingly for all methods the bias in parameter estimates decreased as the dataset size increases. The IGLS algorithm gives estimates (of the level 2 variance) that are biased low for small samples but this is rectified by the RIGLS algorithm. For the variance components model two 'diffuse' priors were considered for the variance parameters with one giving reasonably unbiased posterior median estimates ($\Gamma^{-1}(\varepsilon,\varepsilon)$ prior) and one giving reasonably unbiased posterior mode estimates (Uniform$(0,1/\varepsilon)$ prior) for all but the smallest designs. For the random slopes model Browne and Draper (2000) compared using MCMC with 2 Wishart priors and a Uniform prior for the level 2 variance matrix along with the IGLS and RIGLS methods. Similar results were seen for IGLS and RIGLS as for the variance components model. The Wishart priors exhibited small biases whilst the Uniform prior performed poorly.

In terms of interval coverage for the variance components model both MCMC methods had far better coverage than the IGLS and RIGLS methods using standard estimate ± 2 standard errors intervals, however Browne and Draper (2004) gave several alternative methods that give better intervals using the RIGLS method. For the random slopes regression model the best method for coverage intervals depended on the dataset size and underling model parameters.

So the general conclusions to be drawn from these simulations are that MCMC methods are useful (if slower) alternatives to the IGLS method. For normal response multilevel models there is no great advantage in using MCMC unless you wish to incorporate prior information or wish to exploit the fact that MCMC gives an estimate of the full posterior distribution. This however is useful as we can also calculate full posterior estimates of derived quantities such as residual ranks or the intra-class correlation coefficient which would be difficult using IGLS.

Multilevel Logistic Regression Models

One way of extending the family of models considered so far is to allow response types other than Normal. An example is a multilevel logistic regression model:

$$y_{ij} \sim \text{Binomial}(1, p_{ij})$$
$$\text{logit}(p_{ij}) = X_{ij}\beta + Z_{ij}u_j, u_j \sim N(0, \sigma_u^2)$$

To fit this model in IGLS is not trivial and one solution is to linearize the model via Taylor series expansion. This approach will then lead to quasi-likelihood estimates (see for example Breslow and Clayton 1993), either marginal quasi-likelihood (MQL) or penalized quasi-likelihood (PQL) depending on how the linearization is performed. Once again we can also add prior distributions to this model and fit it in a Bayesian framework. Again MCMC algorithms will fit such a Bayesian model although this time the full conditional distributions will not have standard form and so an

alternative method such as a hybrid Metropolis-Gibbs sampling approach (see for example Browne and Draper (2004) for details) or Adaptive Rejection sampling (Gilks and Wild 1992) is required.

Browne and Draper (2004) also compared MQL, PQL and MCMC methods on simulations from a 3 level logistic regression dataset studied by Rodriguez and Goldman (1995). They showed that both MQL and PQL gave very biased estimates of the higher level variance parameters and poor coverage properties whilst (due to the large number of level 2 and 3 units) both MCMC priors ($\Gamma^{-1}(\varepsilon,\varepsilon)$ and Uniform(0,1/ε) priors) gave much better performance in terms of both bias and coverage intervals. So multilevel logistic regression models give a first example of models where MCMC methods come into their own and have distinct advantages over the IGLS algorithm.

3 Cross-classified and multiple membership models

Nested random effect models allow reasonably sophisticated underlying data structures to be modelled however not all problems support a nested structure. For example in education we may be interested in modelling both school effects and neighbourhood effects. Now all pupils in one school may not live in the same neighbourhood and all children in one neighbourhood may not go to the same school and so we have a *cross-classified* data structure. Another example from education involves looking at a response that occurs at the end of schooling. Now we may be interested in the impact of the school attended on this response however some pupils will have changed schools during their education and each of these schools will have an impact on their schooling. Such a situation will result in a *multiple membership model* where a response will be affected by a (weighted) group of random effects from a classification.

These two extensions to multilevel models allow the user to fit a far greater family of statistical models. We will combine these two extensions into a general framework with a new notation and supporting diagrams later in this section but first we will highlight how the IGLS method can be adapted to fit such models.

IGLS Estimation

As we stated earlier the IGLS algorithm is designed to fit general multivariate normal models of the form $Y \sim MVN(X\beta, V)$ and we can in fact write both a cross-classified and a multiple membership model in this form. However the main reason the IGLS algorithm is feasible for nested models is the fact that the variance matrix V is block diagonal and hence easy to invert. This is not true for these two extensions and inverting a general variance matrix becomes impractical for large problems. A solution for cross-classified models with two classifications (Rasbash and Goldstein 1994) is to split the V matrix into two components one of which is block diagonal and one non-block diagonal part. Then the non-block diagonal part can be dealt with by creating a constrained nested model that contains one variance for each random effect in the smaller of the two classifications (see Rasbash and Browne, 2005 for an example). All of these variances are constrained to be equal and then our constrained nested model is equivalent to the original cross-classified model.

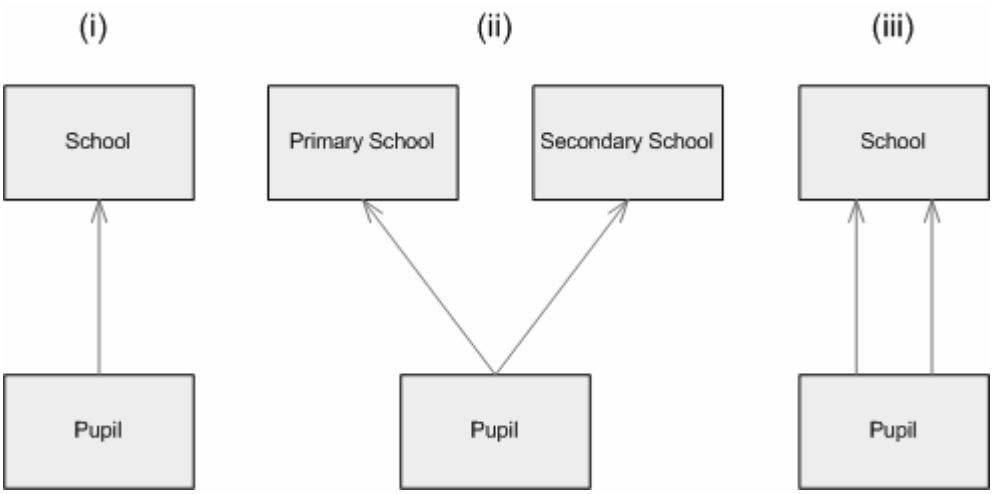
For multiple membership models a similar conversion is available as shown in Rasbash and Goldstein (1994). This method of extension by conversion to a constrained nested model is certainly mathematically elegant but it is limited in the size of model that it can feasibly be used on. Also

extensions to random coefficients or further crossed classifications increase the number of constraints that need including. It is however very useful for small scale problems.

MMMC Models and MCMC Estimation

Cross-classified models and multiple membership models are members of a larger class of models described by Browne et al. (2001) as multiple membership multiple classification (MMMC) models. Here we allow models with many classifications which can be either standard or multiple membership classifications and can be either nested or crossed with each other. Browne et al. (2001) introduced the classification diagram as shown in Figure 1 which allows a pictorial representation of the underlying data structure and a notation (using classification identifiers as indexes) that is easy to extend to such models.

Figure 1: Classification Diagrams for
(i) a nested model (ii) a cross-classified model and (iii) a multiple membership model



To illustrate the notation if we consider an MMMC model for an educational dataset with two higher level classifications, primary school and secondary school where secondary school is a multiple membership classification as we have information on all the secondary schools a child has extended. Note this model is a combination of the models (ii) and (iii) in Figure 1 and its classification diagram would look similar to (ii) but with a double arrow pointing to the secondary school box. Then we write this in the notation of Browne et al. (2001) as

$$y_i = X_i\beta + Z_i^{(2)}u^{(2)}_{primary(i)} + \sum_{j\in \mathrm{sec}ondary(i)} w^{(3)}_{i,j} Z_i^{(3)} u_j^{(3)} + e_i$$

$$u_i^{(2)} \sim N(0,\Sigma_{u(2)}), u_i^{(3)} \sim N(0,\Sigma_{u(3)}), e_i \sim N(0,\sigma_e^2)$$

Here y is an N vector, β is a vector of p_f fixed effect parameters, and $u_i^{(2)}, u_i^{(3)}$ are the vectors of residuals for the p_2 and p_3 random effects for the primary and secondary school classifications respectively. The e_i are scalar pupil level residuals. $X_i, Z_i^{(2)}$ and $Z_i^{(3)}$ are vectors of predictor values and $w_{i,j}^{(3)}$ is a scalar weight for secondary school j for pupil i that represents the proportion of time they spend in the school. This notation is then easily extendable to models with many classifications and Browne et al. (2001) consider three interesting examples: A model for salmonella in Danish poultry

that accounts for both a breeding and production hierarchy, a model for household migration in Belgium that accounts for individual factors and previous households lived in and a model that uses multiple membership effects to account for spatial variation in lip cancer rates in Scotland.

Browne et al. (2001) focuses on MCMC estimation and it is important to realize that for the MCMC algorithm a crossed structure is no more complicated than a nested structure as each classification is treated as an additive term in the model rather than impacting on a global V matrix. This means that the full conditional distributions are similar in form regardless of whether classifications are nested or crossed. In fact the notation used above gives no information on the data structure as it is not needed by the algorithm and this is the motivation for also using a classification diagram when representing the model.

4 Further extensions

In this section we will describe some more recent work that is being considered by researchers associated with the MLwiN software. This work has two main aims, first to (incrementally) extend the family of models available to account for further complications in the data structure and second to provide supporting tools to the existing models that can be fitted in MLwiN. Most of these further extensions deal with cases where we have more than one response variable of interest and are built on an earlier extension to MLwiN that allows multivariate Normal response models using MCMC (see chapter 17 of Browne (2003)).

Multilevel factor models

Often in survey research, in particular when we deal with questionnaires, we have many responses for each individual. Interest often then lies in reducing this multitude of responses by finding interesting underlying latent concepts that are represented by the responses. Techniques like factor analysis are often used to reduce the dimension of questionnaire data to a more manageable set of latent responses. Such single level factor models do not take account for the underlying structure in the population on which the questionnaires are administered. A natural extension is to combine factor analysis modelling with multilevel modelling and this leads to multilevel factor modelling (e.g. Goldstein and Browne 2002). Again it is fairly easy to extend MCMC methods to fit such models although care has to be taken (as with all factor models) to ensure the models are identifiable. This is slightly more of an issue with MCMC methods as there are often multiple symmetric optimal solutions and although a maximum likelihood method will converge to one optimal, MCMC methods may move between solutions.

Goldstein and Browne (2005) show how MCMC can be used to fit multilevel factor models with both binary and continuous responses and further work is now being considered to look at other response types, for example ordered and unordered category responses. One of the great features of MCMC estimation is that new functionality can be ‘bolted on’ to existing models and so often extension in one direction allows lots more model types to be fitted. For example implementing steps to fit factor analysis structures in MLwiN not only allows standard multilevel factor analysis models but also cross-classified factor analysis models without any additional work!

Response variables at different levels

Related to the multilevel factor analysis models, there are many other similar complicated problems that lead to further extensions to the standard multilevel model and these are currently being studied by the Centre for Multilevel Modelling team at the University of Bristol. One possible extension is when we have responses that when combined in a random effect framework, appear at different levels of the model. For example a survey may be administered on households where some responses are measured on the household as a unit, while other responses may be measured on each individual in the household. We would like to fit a multivariate model that accounts for correlations between the household level residuals for the individual level responses and the responses measured at the household level. In fact this is one type of model that the IGLS algorithm can easily accommodate as it is fitting a general multivariate normal model and a nested model with responses at more than one level will even result in a block diagonal V matrix. The adaptation to the standard MCMC algorithm is not difficult and will be considered by the Bristol project.

Missing data and multiple imputation

Missing data is a problem that proliferates in most data collected that we might consider using to fit multilevel models. Both the IGLS algorithm and MCMC can handle missing responses in a multivariate normal response model assuming a missing at random (MAR) mechanism although their techniques are rather different. IGLS fits general multivariate normal models and any individuals who do not have complete responses simply have their missing responses omitted when the response vector y is formed. The V matrix will still be block diagonal (for nested models) but the blocks at the bottom level will differ depending on the missing pattern of responses. MCMC on the other hand treats the missing responses in the model as additional parameters that need estimating. The MCMC algorithm then requires an additional step where the missing responses are generated from their conditional distributions at each iteration.

This links in nicely with multiple imputation (Rubin 1987) which involves producing several complete datasets from a dataset with missing data and then combining analyses performed on each of the complete datasets. To generate complete datasets that respect the multilevel structure of the data we can fit a multivariate response model in MCMC and at every Nth iteration output the complete dataset created by the observed responses and the current estimates for the missing data (as described in chapter 17 of Browne (2003)). Here a large value of N may be required to ensure that the datasets are independent. Of course generating imputation datasets and then fitting the same model to them as was used to create the datasets has no real advantages over simply fitting the model using MCMC. However we can consider constructing an imputation model that is used simply to generate missing values and then a separate model that is our real model of interest. This allows us to generate values for missing observations in both responses and predictors in our model of interest using our imputation model. James Carpenter has built some macros around the MLwiN MCMC engine that will do precisely this for continuous responses. It is hoped that this work will be extended in the future to cover other response types such as categorical data using latent variable ideas.

Sample size calculations

When performing a data collection exercise we generally have some hypotheses that we wish to test from the data we collect. Questions that face researchers are how much data needs collecting and how should that data be collected? If we could collect independently identically distributed (iid) data then there are formulae that will give the sample sizes required to estimate an (assumed) effect size at a required significance level with a particular power. Of course this iid assumption will not be true for

the typical models we have discussed in this paper and so the simple formula cannot be used. Snijders and Bosker (1993) considered the problem of sample size estimation in two level nested models but their approach relies on many factors such as balanced designs. As we have discussed earlier Browne and Draper (2004) performed simulation studies that showed that estimation methods for multilevel models performed better for larger study designs. Their approach involves generating many datasets similar to the desired dataset and looking at the estimates produced, and this approach is also useful for sample size calculations. If we generate datasets similar to those we expect to collect then we can assess if the sample sizes we are considering will allow us to estimate the desired effect size at the required significance level.

The ESRC is funding a project at the University of Nottingham starting in September 2005 that will produce a piece of software to automate such sample size calculations for many of the models considered here including both nested and cross-classified structures. This dataset generating program can then be used in conjunction with any estimation method and given a suitable interface any statistical software package estimation engine to answer sample size questions.

5 Conclusions

In this short article we have attempted to show the flexibility of MCMC methods in attempting to match the complexity of data structures found in real problems. We have contrasted the ease of extension of models when using MCMC with the difficulties in extending the IGLS algorithm. We have attempted to illustrate our incremental approach to this extension where particular applications have influenced our choice of new families of models to consider. We have also illustrated how supporting research is required to supplement these new model families. For example diagrams and notation to represent the new model families and issues such as sample size calculations and missing data that may affect these new models.

It should be noted that in this paper we have strongly focussed on models and methods that have been considered in development of the MLwiN software package. There are many other fine software programs that fit similar models and use other methods. In particular WinBUGS is a package that also uses MCMC estimation and although it is typically slower for fitting the same model it offers a far more extensive range of models. The Genstat package is a great alternative REML based package that is very good at fitting cross-classified models. The GLLAMM package in Stata is a useful alternative for binary response models as it offers maximum likelihood estimation (through quadrature estimation) although it can be slow. The Centre for Multilevel Modelling has carried out an extensive review of software for fitting many of the models described in this paper and the interested reader is recommended to look at http://multilevel.ioe.ac.uk/softrev/index.html

References

Breslow, N.E. and Clayton D.G. (1993)
Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Browne, W.J. (2003)
MCMC Estimation in MLwiN. London: Institute of Education, University of London.

Browne, W.J. and Draper, D. (2000)
Implementation and performance issues in the Bayesian fitting of multilevel models. *Computational Statistics*, **15**, 391-420.

Browne, W.J. and Draper, D. (2004)
A Comparison of Bayesian and Likelihood-based methods for fitting multilevel models. *University of Nottingham Research Report* 04-01

Browne, W.J., Goldstein, H. and Rasbash, J. (2001)
Multiple membership multiple classification (MMMC) models. *Statistical Modelling* **1**: 103-124.

Gelfand, A.E. and Smith, A.F.M. (1990)
Sampling Based Approaches to calculate marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gilks, W.R. and Wild, P. (1992)
Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337-348.

Goldstein, H. (1986)
Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43-56.

Goldstein, H. (1989)
Restricted unbiased iterative generalized least squares estimation. *Biometrika*, **76**, 622-623.

Goldstein, H. and Browne, W. J. (2002)
Multilevel factor analysis modelling using Markov Chain Monte Carlo (MCMC) estimation. In Marcoulides and Moustaki (Eds.), *Latent Variable and Latent Structure Models*. p 225-243. Lawrence Erlbaum, New Jersey.

Goldstein, H. and Browne, W.J. (2005)
Multilevel Factor Analysis Models for Continuous and Discrete Data. In Maydeu-Olivares, A and McArdle, J.J. (Eds.), *Contemporary psychometrics: a festschrift for Roderick P. McDonald*, p 453-475. Lawrence Erlbaum, New Jersey.

Rasbash J. and Browne W. J. (2005)
Non-Hierarchical Multilevel Models. To appear in De Leeuw, J. and Kreft, I.G.G. (Eds.), *Handbook of Quantitative Multilevel Analysis.*

Rasbash, J. and Goldstein, H. (1994)
Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. *Journal of Behavioural Statistics* **19**, 337-350.

Rasbash, J., Steele, F., Browne, W.J., and Prosser, B. (2004)
A User's Guide to MLwiN. Version 2.0: Institute of Education, London.

Rodriguez, G. and Goldman, N. (1995)
An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73-89.

Rubin, D.B. (1987)
Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley and Sons.

Snijders, T.A.B. and Bosker, R.J. (1993)
Standard Errors and Sample Sizes for 2-level research. *Journal of Educational Statistics*, **18**, 237-259

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2000)
WinBUGS version 1.3: user manual. Cambridge: Medical Research Council Biostatistics Unit.

About the Author

William Browne (william.browne@nottingham.ac.uk) is a lecturer in the School of Mathematical Sciences, University of Nottingham.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Small Area Estimation under a Two-Part Random Effects Model with Application to Estimation of Literacy in Developing Countries

Danny Pfeffermann, Bénédicte Terryn, Fernando Moura

Abstract

The UNESCO Institute for Statistics has initiated a programme to collect data on the literacy skills of adults in developing countries. This involves conducting small-scale surveys in a few countries, which consist of administering interviewees aged 15+ a test to measure their literacy score. One objective of this programme is to obtain summary measures of literacy levels in geographical areas for which only very small samples would be available, thus requiring the use of model based small area estimation methods.

Available methods are not suitable, however, for this kind of data due to the mixed distribution of the literacy scores in developing countries. This distribution has a large peak at zero, i.e., a large proportion of adults that are illiterate, and juxtaposed to this peak is an approximately bell-shaped distribution of the non-zero scores measured for the rest of the sample.

In this presentation we will develop a two-part three-level model that is suitable for this kind of data and show how to obtain the small area measures and their variances, or compute confidence intervals, based on this model. The proposed method will be illustrated using simulated data and data obtained from a literacy survey conducted in Cambodia.

Keywords

MCMC, generalized linear mixed model, linear mixed model

1 Introduction

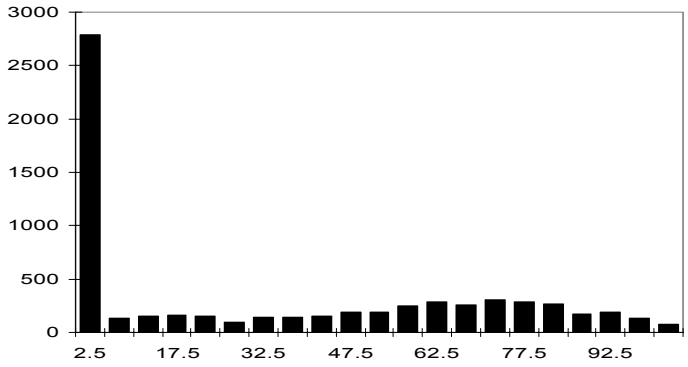
We consider the distribution of scores obtained from literacy tests administered to adults in a household survey. In most developing countries, where many people cannot read or write, this is not a standard distribution. Typically, it consists of a large peak at zero, juxtaposed to a continuous distribution for the non-zero scores, as observed, for example, in a literacy survey carried out in Cambodia in 1999 (see Figure 1 below).

In this paper we discuss ways of producing literacy estimates for areas with small samples under such a mixed distribution. This kind of mixed distribution has not been considered before in the small area estimation literature. The proposed model consists of two parts. The first part is a logistic model used to predict the probability of a positive score. The second part is a standard model (linear model with normal error terms in our application) used to predict a non-zero score. Both models include individual and area level covariates as well as random effects that account for variations not explained by the

covariates. The model accounts for correlations between the corresponding random effects of the two parts. The model is fitted by application of Markov Chain Monte Carlo (MCMC) simulations.

The two-part model is applied to data collected as part of a national literacy household survey carried out in Cambodia in 1999, known as the 'Assessment of the Functional Literacy Levels of the Adult Population'. The performance of the proposed model is tested by simulating data sets that mimic the Cambodia data. The use of simulations also enables us to compare the results of fitting the full model with the results obtained when fitting the two parts of the model separately, without accounting for the correlations between the random effects in the two parts. Another comparison of interest is to results obtained when ignoring the special nature of the data and fitting the linear part to the whole data, ignoring the problem of many zero scores.

Figure 1. Histogram of literacy scores in a national literacy survey in Cambodia, 1999



2 Model and small area predictors

Let *Y* define the response value (literacy test score in our application) and *R* the covariate variables and random effects. Then,

$$+E(Y \mid R=r, Y>0)\Pr(Y>0 \mid R=r) = E(Y \mid R=r, Y>0)\Pr(Y>0 \mid R=r) \quad (1)$$

since $E(Y \mid R=r, Y=0)=0$. For the small area estimation application considered in this paper we consider a nested 3 level model with districts of residence defining the first level, villages defining the second level and individuals defining the third level. For individual *k* residing in village *j* of district i, we have therefore the relationship,

$$E(y_{ijk} \mid r_{ijk}=r) = E(y_{ijk} \mid r_{ijk}=r, y_{ijk}>0)\Pr(y_{ijk}>0 \mid r_{ijk}=r) \quad (2)$$

In what follows we model the two parts in the right hand side of (2). For individuals with positive responses we assume the familiar 'linear mixed model',

$$y_{ijk} \mid r_{ijk}, y_{ijk}>0 = x_{ijk}'\beta + u_i + v_{ij} + \varepsilon_{ijk};\ u_i \sim N(0,\sigma_u^2)\ ;\ v_{ij} \sim N(0,\sigma_v^2)\ ;\ \varepsilon_{ijk} \sim N(0,\sigma_\varepsilon^2) \quad (3)$$

where x_{ijk} represents individual and area level values of the covariates, u_i is a random district effect and v_{ij} is a nested random village effect. The random effects and the residual terms ε_{ijk} are assumed to be mutually independent. Notice that by (3),

$$E(y_{ijk} \mid r_{ijk}, y_{ijk}>0) = x_{ijk}'\beta + u_i + v_{ij} \quad (3')$$

The random effects account for the variation of the individual responses not explained by the covariates. Alternatively, they define the correlations holding between the responses of individuals residing in the same village, or individuals residing in the same district but in different villages.

$$Corr(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} (\sigma_u^2 + \sigma_v^2)/(\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2) & \text{if } j = j', k \neq k' \\ \sigma_u^2/(\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2) & \text{if } i = i', j \neq j' \\ 0 & \text{if } i \neq i' \end{cases} \quad (4)$$

For the probabilities of positive responses (second part of (2)) we assume the 'generalized linear mixed model',

$$\Pr(Y_{ijk} > 0 | x_{ijk}, u_i^*, v_{ij}^*) = p_{ijk} = \frac{\exp(x_{ijk}'\gamma + u_i^* + v_{ij}^*)}{1 + \exp(x_{ijk}'\gamma + u_i^* + v_{ij}^*)};\ u_i^* \sim N(0, \sigma_{u^*}^2)\ ;\ v_{ij}^* \sim N(0, \sigma_{v^*}^2) \quad (5)$$

implying, $logit(p_{ijk}) = \log \frac{p_{ijk}}{1 - p_{ijk}} = x_{ijk}'\gamma + u_i^* + v_{ij}^*$. Here again u_i^* and v_{ij}^* represent random district and village effects not accounted for by the covariates.

The proposed model permits nonzero correlations between the district random effects in the two parts, and similarly for the village random effects. This is a reasonable assumption since it can be expected that for given values of the covariates, an individual residing in an area characterized by high literacy scores will have a higher probability of a positive score than an individual residing in an area with low scores. See Figures 2 and 3 below for some supporting evidence from data in Cambodia. (The correlations are 0.35 for villages and 0.38 for districts.) The correlations are modelled by assuming,

$$u_i^* | u_i \sim N(K_u u_i, \sigma_{u^*|u}^2)\ ;\ v_{ij}^* | v_{ij} \sim N(K_v v_{ij}, \sigma_{v^*|v}^2) \quad (6)$$

Let U_i define the population of first level i of size N_i. The small area parameters of interest are the means, $\overline{Y}_i = \sum_{j,k \in U_i} y_{ijk} / N_i$, which in the case of the survey in Cambodia are the true district means of the literacy scores. Notice that the means are computed over all the individuals in the area, including individuals with zero scores. Under the model defined by (2), the means can be predicted as,

$$\hat{\overline{Y}}_i = \frac{1}{N_i}\{\sum_{j,k \in s_i} y_{ijk} + \sum_{j,k \notin s_i} [\hat{E}(Y_{ijk} | r_{ijk}, Y_{ijk} > 0) \times \hat{p}_{ijk}]\} \quad (7)$$

where s_i defines the sample from first level (district) i. By (3) and (5), the predictor in (7) takes the form,

$$\hat{\overline{Y}}_i = \frac{1}{N_i}[\sum_{j,k \in s_i} y_{ijk} + \sum_{j,k \notin s_i} (x_{ijk}\hat{\beta} + \hat{u}_i + \hat{v}_{ij}) \times \frac{\exp(x_{ijk}'\hat{\gamma} + \hat{u}_i^* + \hat{v}_{ij}^*)}{1 + \exp(x_{ijk}'\hat{\gamma} + \hat{u}_i^* + \hat{v}_{ij}^*)}] \quad (8)$$

with $\hat{\beta}, \hat{\gamma}, \hat{u}_i, \hat{v}_i, \hat{u}_i^*, \hat{v}_{ij}^*$ defining appropriate sample estimates (see next section).

Figure 2. Proportion literate by average score for districts in center of Cambodia, 1999 survey

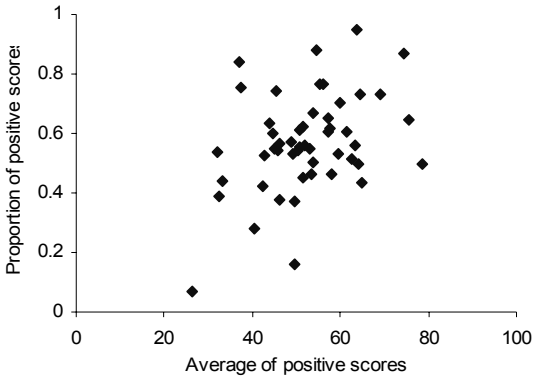
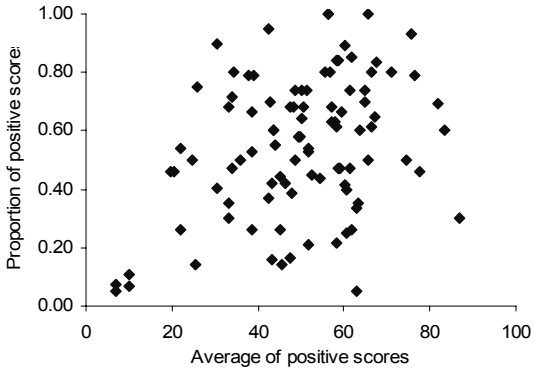


Figure 3. Proportion literate by average score for villages in center of Cambodia, 1999 survey



3 Inference

The use of the small area predictors defined by (8) requires estimating the fixed parameters $(\beta,\sigma_u^2,\sigma_v^2,\sigma_\varepsilon^2)$ of the linear part (Equation 3), the fixed parameters $(\gamma,K_u,K_v,\sigma_{u^*|u}^2,\sigma_{v^*|v}^2)$ of the logistic part (Equations 5, 6), and predicting the random effects $\lambda_{ij}=\{(u_i,v_{ij};u_i^*,v_{ij}^*)\}$. Methods for estimating the fixed and random effects when fitting linear mixed models or generalized linear mixed models alone have been developed over the last two decades under both the frequentist and the Bayesian paradigms. The use of these methods permits also the computation of estimators of the mean square error (MSE) of the small area predictors that account for parameter estimation to correct order; see the recent book of Rao (2003) for a thorough review and discussion. However, the two-part model defined by (2)-(6) has not been considered in the literature in the context of small area estimation and in what follows we describe briefly a few possibilities of fitting this model.

Likelihood based inference

Define, $\mathrm{I}_{ijk}=1$ if $Y_{ijk}>0$, $\mathrm{I}_{ijk}=0$ if $Y_{ijk}=0$, and denote $r_{ijk}=(x_{ijk},u_i,v_{ij})$, $r_{ijk}^*=(x_{ijk},u_i^*,v_{ij}^*)$. For given vectors r_{ijk}, r_{ijk}^*, the likelihood for the two-part model takes the form,

$$L=\prod_{i,j,k\in s}(p_{ijk})^{\mathrm{I}_{ijk}}f(y_{ijk}\mid y_{ijk}>0)^{\mathrm{I}_{ijk}}(1-p_{ijk})^{(1-\mathrm{I}_{ijk})} \qquad (9)$$

with p_{ijk} and $f(y_{ijk}\mid y_{ijk}>0)$ defined by (5) and (3) respectively and $s=\cup s_i$ denoting the sample from all the areas. The use of this likelihood for inference is, however, problematic because the random effects $\lambda_{ij}=\{(u_i,v_{ij};u_i^*,v_{ij}^*)\}$ are in fact unobservable random variables. One possibility, therefore, is to integrate the likelihood over the joint (normal) distribution of the random effects as defined by (3) and (6), and maximize the integrated likelihood with respect to the fixed parameters $(\beta,\sigma_u^2,\sigma_v^2,\sigma_\varepsilon^2)$ and $(\gamma,K_u,K_v,\sigma_{u^*|u}^2,\sigma_{v^*|v}^2)$. The integrated likelihood is

$L_0 = \int \prod_{i,j,k \in s} (p_{ijk})^{I_{ijk}} f(y_{ijk} \mid y_{ijk} > 0)^{I_{ijk}} (1 - p_{ijk})^{(1-I_{ijk})} g(\lambda_{ij}) d\lambda_{ij}$. Olsen and Schafer (2001) consider a similar two-part model for fitting longitudinal data. The authors approximate the integrated likelihood by a high order multivariate Laplace approximation (Raudenbush *et al*. 2000), and calculate empirical Bayes predictors of the random effects by use of importance sampling (Tanner, 1996), setting the fixed parameters at their maximum likelihood estimates. The application of this procedure, however, is complicated computationally and the MSE estimators of the small area predictors obtained this way fail to account for the variation induced by estimating the fixed parameters. The contribution of fixed parameter estimation to the total MSE can not be ignored in general, unless the number of sampled areas is large.

Separate model fitting:

The idea here is to fit the two parts of the model separately and then combine the estimates for computing the small area predictors in (8). As mentioned earlier, the fitting of the separate parts has been researched extensively in the literature over the last 2 decades and computer software is now readily available, particularly for linear mixed models. See Rao (2003) for a short review. Note that extra care should be taken when computing MSE estimators of correct order.

The use of separate model fitting has, however, the disadvantage of not lending itself to simple computations of the coefficients (K_u, K_v) (Equation 6), and it is not clear how to account for the existing correlations between the two data sets for enhancing the efficiency of the small area predictors. Notice also that by ignoring the correlations between the random effects of the two parts of the model, the estimated MSE of the small area predictors are imprecise.

Bayesian inference

The use of Bayesian methods requires specification of prior distributions for all the fixed parameters underlying the two-part model, but with the aid of Markov Chain Monte Carlo (MCMC) simulations the use of this approach permits sampling from the posterior distribution of the fixed parameters and the random effects, and hence from the posterior distribution of the small area means, given the data. Thus, the MCMC algorithm yields the whole posterior distribution of the small area means of interest, and hence correct MSE (posterior variance) estimators or confidence (credibility) intervals can be computed. Computer software is available to perform all the necessary computations but it should be noted that with complex models the computations can be very intensive and time consuming.

For the empirical study of this article we followed the Bayesian approach using the WinBUGS software (Spiegelhalter *et al.* 2003), which implements the MCMC algorithm with the Gibbs sampler (Gelfand and Smith, 1990). The Gibbs sampler samples alternately from the conditional posterior distribution of each of the fixed and random parameters (random effects), given the data and the remaining parameters. It defines a Markov chain which, under some regularity conditions converges to the joint posterior distribution of all the model parameters. Thus, at the end of the sampling process (upon convergence), the algorithm produces a (single) realization of each of the fixed and random parameters from their corresponding posterior distribution given the data, and hence a single realization from the posterior distribution of each small area value

$\theta_i = \frac{1}{N_i}[\sum_{j,k\in s_i} y_{ijk} + \sum_{j,k\notin s_i} (x_{ijk}\beta + u_i + v_{ij}) \times \frac{\exp(x_{ijk}'\gamma + u_i^* + v_{ij}^*)}{1+\exp(x_{ijk}'\gamma + u_i^* + v_{ij}^*)}]$ (compare with (8)). Repeating the same sampling process independently a large number of times yields an approximation to the posterior distribution of each of the values θ_i. The true small area mean, $\bar{Y}_i$, can then be predicted by averaging the sampled values θ_i in all the chains. (The average estimates the posterior expectation of the small area mean, see also the comment below.)

The MSE is estimated by computing the empirical variance of the sampled values from the posterior distribution of the means $\bar{Y}_i$. The sampled values are obtained by first predicting the individual nonsampled measurements y_{ijk} and then computing the means $\hat{\bar{Y}}_i = \frac{1}{N_i}[\sum_{j,k\in s_i} y_{ijk} + \sum_{j,k\notin s_i} \hat{y}_{ijk}]$. The predictors $\hat{y}_{ijk}$ are computed by randomly drawing from their posterior distribution, i.e.,

$$\hat{y}_{ijk} = (x_{ijk}\tilde{\beta} + \tilde{u}_i + \tilde{v}_{ij} + \tilde{\varepsilon}_{ijk}) \times \tilde{\mathrm{I}}_{ijk} \qquad (10)$$

where, $\tilde{\mathrm{I}}_{ijk} = 1$ with probability $\tilde{p}_{ijk}$ and $\tilde{\mathrm{I}}_{ijk} = 0$ with probability $(1-\tilde{p}_{ijk})$, $\tilde{p}_{ijk} = \frac{\exp(x_{ijk}'\tilde{\gamma} + \tilde{u}_i^* + \tilde{v}_{ij}^*)}{1+\exp(x_{ijk}'\tilde{\gamma} + \tilde{u}_i^* + \tilde{v}_{ij}^*)}$. Notice that each of the fixed and random effects used for the prediction of the measurements y_{ijk} (denoted by "~") is a random draw from its posterior distribution. Confidence (credibility) intervals with coverage rates of $(1-\alpha)$ are defined by the $\alpha/2$ and $(1-\alpha/2)$ level quantiles of the empirical posterior distribution of the $\bar{Y}_i$ (the distribution of the sampled values $\hat{\bar{Y}}_i$).

In practice, the use of parallel chains for producing independent realizations is often too time consuming, in which case the samples can be generated from a single long chain or a few chains, but selecting only every rth sampled value (after convergence), thus reducing as much as possible the dependencies existing between adjacent sampled values.

Comment: The posterior mean of $\bar{Y}_i$ could also be estimated by simply averaging the sampled values $\hat{\bar{Y}}_i$ from its posterior distribution. Notice, however, that these values contain also the sums $\sum_{j,k\notin s_i} \hat{\varepsilon}_{ijk}/N_i$ for which the posterior mean is zero, so that the use of this procedure adds some extra noise to the estimation of the posterior mean if the number of MCMC simulations is not sufficiently large (depending also on the posterior variance of the ε_{ijk}).

4 Empirical Results

We use data from the 1999 survey, 'Assessment of the Functional Literacy Levels of the Adult Population' in Cambodia for the empirical illustrations. This is a household survey that had 6548 adults being interviewed and administered a literacy test consisting of 20 tasks in the Khmer language, with scores ranging from 0 to 100 (see Figure 1 in the introduction). It used a stratified multi-stage sampling design with the strata defined by the 24 provinces that comprise the country. Within each of

the provinces half of the districts were selected, then within each district 2 communes were selected and within each commune, 3 villages were selected (with a few exceptions). Finally, households were selected in each village and one adult sampled from each household, altering according to age and sex. The sampling scheme at each stage was systematic sampling. The number of households selected in each village was constant for all the sampled villages belonging to the same province. The province total sample sizes were allocated proportionally to the population province sizes.

In what follows the small areas of interest are the country districts, with sample sizes varying between 0 (no sample) in 88 of the districts to almost 150 in the districts of the capital city. Twenty one districts had sample sizes less than 20, and another 16 districts had sample sizes between 41 and 60. The data analyzed for this study refer to the 50 rural districts in provinces located in the center of the country. The total sample size is n=4028.

Table 1 shows the results obtained when fitting the logistic model alone to this data set, with and without random effects for districts and villages. The dependent variable I_{ijk} takes the value 1 if $y_{ijk} > 0$ and takes the value 0 otherwise, see Equation (5). Table 2 shows the results of fitting the linear model to individuals with positive scores alone, again with and without the inclusion of random effects. These two models have been fitted using the MLwiN software (Goldstein, 2003). This software computes maximum likelihood estimators of the fixed parameters and empirical best linear unbiased predictors (EBLUP) of the random effects for linear mixed models, and predictive quasi likelihood estimators (PQL) of the fixed parameters and random effects for generalized linear mixed models. (Other estimation procedures are also available.) The regressor variables in the two models have been chosen by application of some standard model selection procedures, without the inclusion of the random effects. All the variables except those referring to age, education and household size are dummy variables taking the value 1 when the variable definition is satisfied.

Table 1. Model parameters and standard errors (S.E.) when fitting logistic part alone

	Without random effects		**With random effects**	
Variables	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Constant	-4.80	0.44	-6.48	0.58
No school, attended literacy prog.	2.07	0.21	2.44	0.27
Education	1.75	0.09	2.16	0.12
Education^2	-0.11	0.01	-0.13	0.01
Helped by interviewer	1.09	0.11	2.00	0.17
Living in a remote area	-0.56	0.21	-0.32	0.49$^{(*)}$
Gender (1 for female)	-0.63	0.11	-0.59	0.14
Having low income	-0.39	0.11	-0.35	0.14
Age	0.11	0.02	0.14	0.02
Age^2	-0.001	0.000	-0.002	0.000
Random effects			**Variance**	**S.E.**
Between district			1.28	0.34
Between villages			0.86	0.19

(*) not significant

Table 2. Model parameters and standard errors (S.E.) when fitting linear part alone

	Without random effects		**With random effects**	
Variables	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Constant	5.85	4.18	6.90	4.00
Civil servant and professional	10.71	2.14	13.91	1.89
Education	6.64	0.59	7.28	0.53
Education^2	-0.19	0.05	-0.24	0.05
Low income	-3.15	0.94	-2.61	0.88
Gender (1 for female)	-2.52	0.92	-1.60	0.81
Number of adults in household	1.23	0.31	0.94	0.29
Age	1.04	0.18	0.84	0.16
Age^2	-0.013	0.002	-0.010	0.002
Random effects			**Variance**	**S.E.**
Between district			66.31	16.72
Between villages			66.58	10.45
Individual level			322.03	10.12

The main results emerging from the two tables can be summarized as follows: inclusion of the random effects in the two models changes the values of the coefficients, more so in the linear part, but not to the extent of changing their signs. The variances of the random effects when included in the model are highly significant, indicating their contribution to explaining the variation of the scores. Finally, we note the interesting outcome that in the logistic case the standard errors of the estimated coefficients when including the random effects in the model are always larger or equal than the corresponding standard errors when fitting the model without them, and that it is the other way around in the linear case. We don't have a clear explanation to this outcome.

How well do the models fit the data? As noticed from the two tables, all the coefficients except for one in Table 1 are significant (based on standard t-tests), with and without the inclusion of the random effects, and likewise the variances of the random effects. Other variables considered for inclusion in the two models were found to be nonsignificant. The value of R-square for the linear model without the random effects is 0.302. As a further diagnostic for the logistic model we show in Figure 4 a scatter plot of the observed proportions of 'ones' (positive scores) against the average of the predicted probabilities of ones in groups of 50 individuals defined by the ordered values of the predicted probabilities. The predicted probabilities, $\hat{p}_{ijk}$, were computed under the model with random effects. The plotted values are almost on a straight line, showing a good fit. Figure 5 shows a histogram of the estimated standardized individual errors, $\hat{z}_{ijk} = \hat{\varepsilon}_{ijk} / SD(\hat{\varepsilon}_{ijk}) = (y_{ijk} - x_{ijk}'\hat{\beta} - \hat{u}_i - \hat{v}_{ij}) / SD(\hat{\varepsilon}_{ijk})$, when fitting the linear model with random effects to the individuals with scores $y_{ijk} > 0$. Although not a 'perfect' bell shape, the histogram does not signal severe divergence from a normal distribution.

Figure 4. Observed and predicted probabilities of positive scores, logistic model

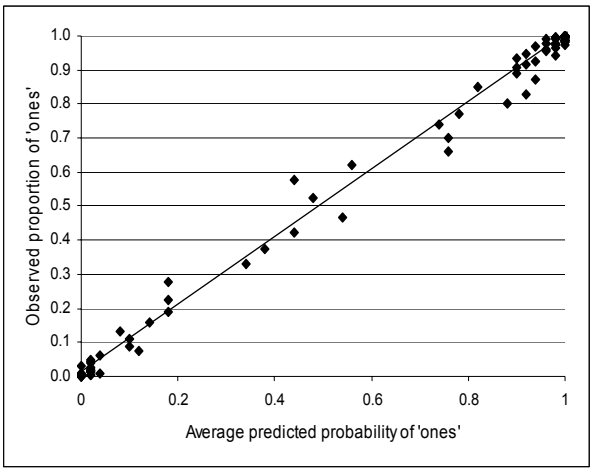
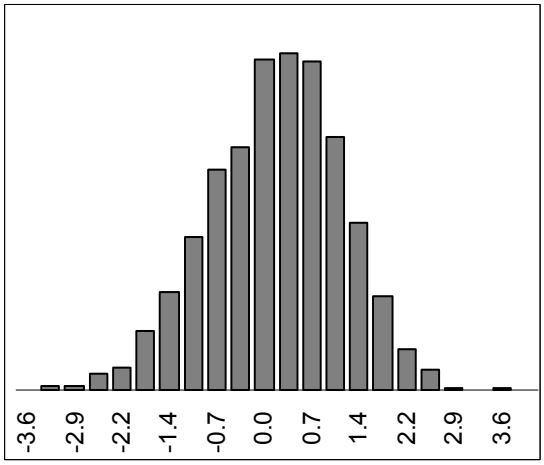


Figure 5. Histogram of standardized residuals, linear model



5 Simulation Study

The purpose of the simulation experiment is to study the effectiveness of the two-part model for producing small area predictors and measures of the associated prediction errors, and to compare the results with results obtained when fitting the two parts of the model separately ignoring the correlations between the corresponding random effects in the two parts, and with the results obtained when fitting a linear mixed model to the whole sample ignoring the accumulation of zero scores. To this end, we generated 300 populations of size 4,028 (the same size as the original data set analyzed in Section 4), and 300 corresponding samples of size 750, using a two-part model but with fewer regressors than in Tables 1 and 2. In the logistic part we included 4 regressors: 'number of years at school', 'attendance of a literacy programme', 'helped by the interviewer' and 'having low income'. In the linear part we included 5 regressors: 'number of years at school', 'gender', 'number of adults in the household', 'age', and 'age^2'. In order to specify sampled values of the regressor variables, we sampled at random 750 individuals from the data set considered in Section 4, and the observed regressor values were then used for all the 300 samples. New random effects and errors were generated for every simulation using the model defined by (3) and (6), and added to the fixed effects $x_{ijk}'\gamma$ and $x_{ijk}'\beta$ in the logistic and the linear parts for every sampled and nonsampled population unit. The correlations between the random effects of the logistic and the linear parts were set to 0.26 for the district random effects and 0.41 for the village random effects. See Table 3 for the other parameter values used for generating the data. Individual scores y_{ijk} were generated by performing Bernoulli trials with probabilities $\Pr(\mathrm{I}_{ijk}=1)=p_{ijk}$ as defined by the logistic model in (5), and in the case of a 'one', generating a score from the model (3). The district means of the y-values in the population (zero and nonzero scores) were taken as the true district means. The samples contained individuals from all the 50 districts, with 11 districts having samples of size $1 \le n_d \le 10$, 29 districts having samples of size $11 \le n_d \le 20$, and the remaining 10 districts having samples of size $21 \le n_d \le 30$. The use of Bayesian estimators requires specifying prior distributions for all the hyper-parameters. We used normal priors with large variances for the elements of the vector

coefficients β, γ, and uniform priors for the standard deviations underlying the two parts of the model and the coefficients K_u and K_v (Equation 6).

We encountered unexpectedly severe computation problems when fitting the two-part model with WinBUGS, accounting for both district random effects and village random effects. The sampled values generated by the Gibbs sampler were found to be strongly correlated even at very high lags, (over 1000 for the village random effects and the correlation between the village random effects in the two parts, and still over 500 when tightening the prior distributions), which required extremely long chains to obtain sufficient data for inference. We also couldn't verify convergence of some of the posterior distributions. This made it impossible to perform a full scale simulation study and we therefore fitted the two-part model with only district level effects, despite generating the data with village random effects as well. (The predictors of the linear part remain unbiased even when ignoring the village effects. This is not true for the logistic part but the bias is small.) We are presently investigating ways of overcoming these computational difficulties. For fitting the models with the district random effects we generated chains of length 20,000, discarded the first 10,000 sampled values as "burn in", and then thinned the chains by taking every 20th sample. This resulted in having 500 sampled values from the posterior distribution of each of the fixed and random parameter values and hence 500 sampled values from the posterior distribution of each of the district means.

The results of the simulation study are shown in Table 3 and Figures 4-6. Table 3 shows the mean estimates of the model parameters and the standard deviations of the means over the 300 simulations, as obtained when fitting three different models to the sample data: A- the two-part model that accounts for the correlation between the district random effects in the two parts (denoted "+ Corr." in the table), B- the two part model that ignores the correlation between the district random effects, i.e., when fitting the two parts of the model separately (denoted "– Corr." in the table), and C- the linear mixed model defined by (3) but fitted to all the *y*-values including the zero scores. This model ignores the accumulation of zero scores but in order to make it more comparable to the fitting of the two part models, we included in this model all the regressors appearing in either the logistic part or the linear part of the two-part model. For comparability reasons we fitted all the three models using the WinBUGS software (thus following the Bayesian paradigm), but it should be noted that fitting the models B and C using MLwiN that is much simpler and faster yields very similar results.

We first discuss the results obtained when fitting the two-part model with or without accounting for the correlation between the district random effects in the two parts (Models A and B). As can be observed from the table, the mean estimates of the regression coefficients in the two parts of the model and the standard deviations of the means are very close under the two models, and the mean estimates are generally close to the corresponding true coefficients, indicating lack of appreciable bias. Note, however, that some of the differences between the mean estimates and the true values are significant, despite being small, which could be explained by the fact that the fitted models do not account for village effects. The estimates of the variances of the random effects are again close under the two two-part models, but they cannot be compared directly to the true values, since the models fitted included only district random effects. Nonetheless, for the linear part the sum of the three true variances and the sum of the two estimated variances under the two models are similar, and the ratio of the variance of the district random effects to the residual variance is likewise preserved. For the logistic part the estimated variance is lower by 12% than the sum of the two true variances. Finally, we mention that the correlation between the district random effects in the two parts of the model is estimated with no bias, but the standard deviation of the estimates is quite high, ($0.01 \times \sqrt{300} = 0.17$).

Table 3. Means and standard deviations (S.D.) of means of estimators of model parameters under three models. 300 simulations

		Simulation mean			Standard deviation of simulation mean		
	True value	Model +Corr.	Model - Corr.	Linear model	Model + Corr.	Model - Corr.	Linear model
Fixed effects - linear part							
Intercept	8.45	9.48	9.97	-14.48	0.44	0.44	0.91
β_1	4.93	4.89	4.85	8.50	0.02	0.02	0.49
β_2	1.26	1.24	1.24	1.07	0.02	0.02	0.06
β_3	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00
β_4	0.59	0.52	0.51	.040	0.04	0.04	0.04
β_5	-1.33	-1.43	-1.43	-2.88	0.11	0.11	0.20
Fixed effects - logistic part							
Intercept	-4.04	-3.90	-3.88		0.03	0.03	
γ_1	1.62	1.58	1.57		0.01	0.01	
γ_2	1.84	1.78	1.77	11.53	0.02	0.02	0.68
γ_3	2.41	2.33	2.31	4.46	0.03	0.03	0.32
γ_4	-0.31	-0.29	-0.29	-1.19	0.02	0.02	0.13
Variances - linear part							
Districts	86.03	95.23	97.41	100.22	1.92	1.92	6.06
Villages	31.85						
Residual	327.28	355.75	354.99	539.29	1.46	1.67	31.20
Variances - logistic part							
Districts	2.09	2.50	2.43		0.05	0.05	
Villages	0.74						
Correlations							
Districts	0.26	0.26			0.01		
Villages	0.41						

Turning to the fitting of the linear mixed model (Model C), the mean estimates of all the coefficients are far away from the true coefficients, which of course is not surprising given that the data were generated from a two-part model, but interestingly enough, the signs of the slope coefficients are preserved.

Figures 6 and 7 show the bias and root mean square error (RMSE) when predicting the true district means under the three models. Let $\hat{\bar{Y}}_d^r$ represent any of the three predictors for a given district *d* as obtained in simulation *r*, and denote by $\bar{Y}_d^r$ the corresponding true district mean. The bias and RMSE are defined as,

$$Bias_d = \sum_{r=1}^{300}(\hat{\bar{Y}}_d^r - \bar{Y}_d^r)/300 \;\; ; \;\; RMSE_d = [\sum_{r=1}^{300}(\hat{\bar{Y}}_d^r - \bar{Y}_d^r)^2/300]^{1/2} \qquad (11)$$

Figure 6. Prediction bias
300 simulations

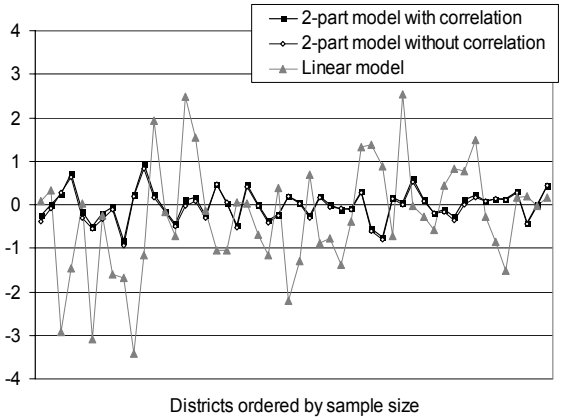
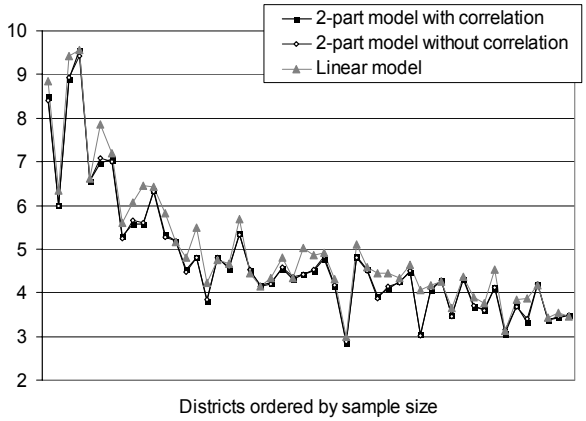


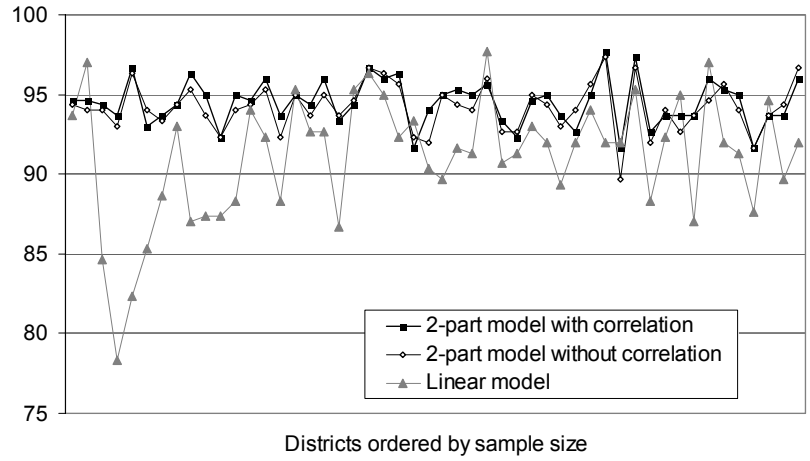
Figure 7. Prediction RMSE
300 simulations



Evidently, fitting the linear mixed model without accounting for the accumulation of zero scores (Model C) yields appreciable biases, irrespective of the sample size. In some districts the absolute biases translate into relative biases of up to 15%. On the other hand, fitting the two-part model A or B, yields basically unbiased predictors. Fitting the linear mixed model yields also larger RMSEs, particularly for districts with small sample sizes, but the increase in RMSE compared to the other two models is not as big as in the case of the bias. This outcome is easily explained by the fact that the variances of the prediction errors are much smaller in the case of the linear mixed model, which is a much simpler model with fewer estimated parameters.

Finally, Figure 8 shows the percentage of times that 95% confidence (credibility) intervals produced under the three models cover the true district means. (See Section 3 for the construction of the confidence interval boundaries when using MCMC simulations.) The prominent result emerging from this Figure is that fitting the linear mixed model ignoring the accumulation of zeroes yields for almost all the districts confidence intervals with lower coverage rates than the nominal 95% rate, with particularly low coverage for districts with small sample sizes. On the other hand, the fittings of the two-part models yield confidence intervals with coverage rates that are close to the nominal 95% rate. In fact, except for one district where fitting the two parts separately yields a coverage rate of 90%, for all the other districts the rates are always between 92% and 97%. There seems to be little difference in the performance of the two two-part models, but we mention that accounting for the correlation between the district random effects in the two parts yields better coverage rates in 28 out of the 50 districts, whereas fitting the two parts separately yields better coverage rates in only 14 districts. In the remaining 8 districts the coverage rates obtained under the two models are the same.

Figure 8. Percentage of confidence intervals covering the true mean



6 Summary

The most important message emerging from this paper is that ignoring the accumulation of zeroes and fitting a linear mixed model can result in biased predictors and undercoverage of confidence intervals. Clearly, the magnitude of the bias and the undercoverage depends on the percentage of zero scores. Fitting a two-part model to such data yields unbiased predictors and confidence intervals with acceptable coverage rates. Fitting the full two-part model, accounting for the correlations between the random effects of the two parts is, in principle, the best choice, but it improved the predictions in our simulation study very marginally, which is probably explained by the low correlation of $\rho_{u,u*} = 0.26$ used for generating the population data.

In this study we used MCMC simulations for fitting the models and computing the small area predictors and their variances, but as mentioned in Section 3, the use of this approach requires specifying prior distributions, which could affect the inference particularly with a small number of sampled areas. The other problem with the use of MCMC simulations is that it is extremely computing intensive. As mentioned earlier, we are presently investigating ways of overcoming the computation problems that we encountered with the use of the WinBUGS program. Another extension of the present study is to fit the full two part model following the frequency approach, using either MLwiN (Goldstein, 2003) or the aML software (Lillard and Panis, 2003).

References

Gelfand, A.E., and Smith, A.F.M (1990)
Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 072-985.

Goldstein, H. (2003)
Multilevel Statistical Models. Third edition. London, Edward Arnold:

Lillard, L.A., and Panis, C.W.A. (2003)
aML Multilevel Multiprocess Statistical Software, Version 2.0. EconWare, Los Angeles, California.

Olsen, M.K., and Schafer, J.L. (2001)
A two-part random effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association*, **96**, 730-745.

Rao, J.N.K. (2003)
Small Area Estimation. New York: Wiley.

Raudenbush, S.W., Yang, M., and Yosef, M. (2000)
Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9**, 141-157.

Spiegelhalter, D., Thomas, A., and Best, N.G. (2003)
Bayesian Inference using Gibbs Sampling. WinBUGS version 1.4, User manual. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K.

Tanner, M. A. (1996)
Tools for Statistical Inference, 3rd edition. New York: Springer - Verlag

About the Authors

Danny Pfeffermann, Hebrew University, Israel, and University of Southampton, U.K.

Bénédicte Terryn, UNESCO Institute for Statistics, Montreal, Canada

Fernando Moura, Federal University of Rio de Janeiro, Brazil

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

EBLUP-type Estimation of Local Authority Unemployment

David Curtis and Ayoub Saei

Abstract

As in many other countries, the Labour Force Survey (LFS) serves as the key source of national information about the UK labour market, and in particular about numbers of unemployed and associated unemployment rates. However, the small sample size of the LFS in many local authority districts (LADs) limits the use of LFS estimates of unemployment at LAD level. Application of standard methods for small area estimation based on linear models also fails in this situation because the response variable of interest (unemployed/not unemployed) is dichotomous. In this paper we describe how an empirical best linear unbiased-type (EBLUP-type) method based on a logistic model for unemployment can be used to estimate unemployment at local areas. This model is an extension of the usual linear logistic model, and includes an LAD-specific random effect in the linear predictor. Estimates of the parameters of the model, including those associated with the random effect, are obtained using maximum likelihood and restricted/residual maximum likelihood methods. In this paper we describe how the Office for National Statistics has implemented this methodology in SAS. We also provide results from a realistic simulation study carried out by the ONS that examines the performance of these EBLUP-type estimators as well as associated estimates of their variability.

Keywords

EBLUP, Logistic regression, Labour Force Survey, REML

1 Introduction

The Labour Force Survey (LFS) serves as the key source of national information about the UK labour market, and in particular about numbers of unemployed and associated unemployment rates. However, the small sample size of the LFS in many local authority districts (LADs) limits the use of LFS estimates of unemployment at LAD level. Application of standard methods for small area estimation based on linear models also fails in this situation because the response variable of interest (unemployed/not unemployed) is dichotomous.

The Labour Force Survey is a continuous, large-scale survey, with a sample of around 60,000 households in each three-month period. These include around 150,000 people, of whom over 110,000 are aged 16 or over, in each three-month period. Since 2000 the Labour Force Survey sample has also included a boost [2]. These data are used to measure unemployment using the International Labour Organisation (ILO) definition on an annual basis covering the period March to February. The sample size within an individual LAD is often too small to provide reliable estimates, typically less than a

quarter of the annual direct LFS estimates of unemployment have relative standard errors under 20% leaving many local authorities with estimates that are not precise enough to base policy decisions on. More information about the Labour Force Survey may be found on the National Statistics web site, in particular see [2] and [3].

The need for a reliable estimate of unemployment and rates at LAD levels is a well known problem at ONS. This paper is an extension of the standard small area estimation method [1] for unemployment and rates at ONS. The linear predictor includes an LAD random effect in addition to the auxiliary information. It is the inclusion of the random effects that differentiates this new methodology from the previous work (standard small area estimation at ONS). The model parameters are estimated by maximum likelihood (ML) and residual/restricted maximum likelihood (REML) method. Section 2 explains the Great Britain LFS 1999 – 2000 data and model. The application and simulation study results are given in sections 3 and 4. The last section is a discussion on the paper.

2 Model and Estimation

Let $\mathbf{y} = \{y_{di}\} = (y_{11}, y_{12}, ..., y_{1I}, ..., y_{d1}, y_{d2}, ..., y_{dI}, ..., y_{D1}, y_{D2}, ..., y_{DI})'$ denotes vector of sample values (LFS data) of the binomial survey variable of interest Y. Assume that LAD random effect u_d is a normal variable with zero mean vector and variance-covariance of φ. Let p_{di} be the probability that an individual in group (age-sex) i from area d unemployed and let n_{di} be the number of individuals in group i from area d in the sample. The model underlying small area estimation for the binomial response data is called logistic linear mixed,

$$\eta_{di} = \text{Logit}(p_{di}) = \ln\left(\frac{p_{di}}{1-p_{di}}\right) = \mathbf{x}'_{di}\boldsymbol{\beta} + u_d$$

where $\mathbf{x}_{di}$ is a vector of group indicator and covariates in the sample and u_d is area random effect follows a normal distribution with zero mean and variance of φ. The models can be rewritten in matrix format as

$$(1) \qquad \boldsymbol{\eta} = \text{Logit}(\mathbf{P}) = \ln\left(\frac{\mathbf{P}}{\mathbf{1}-\mathbf{P}}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where $\mathbf{X}$ is matrix of $\mathbf{x}_{di}$, $\mathbf{1}$ is a vector of all elements 1 and $\mathbf{Z}$ is a known matrix; $\boldsymbol{\eta}$ and $\mathbf{P}$ are

$$\boldsymbol{\eta} = \{\eta_{di}\} = (\eta_{11}, \eta_{12}, ..., \eta_{1I}, ..., \eta_{d1}, \eta_{d2}, ..., \eta_{dI}, ..., \eta_{D1}, \eta_{D2}, ..., \eta_{DI})'$$

$$\mathbf{P} = \{p_{di}\} = (p_{11}, p_{12}, ..., p_{1I}, ..., p_{d1}, p_{d2}, ..., p_{dI}, ..., p_{D1}, p_{D2}, ..., p_{DI})'.$$

There are 408 LADs in Great Britain, 406 are used in this study, i.e., $D = 406$. The City of London and Isles of Scilly are excluded. The key auxiliary variable was claimant count averaged over the same 12 month period (March – February) as the LFS data. In addition to the claimant count, there are age, sex, Government Office Region (12 regions) and ONS area classification (7 categories). The response variable of interest is the number of unemployed. A total of 168,978 people were in the sample. The $\mathbf{X}$ matrix includes the following covariates: -

the logit of the claimant count proportion in each age/sex class within each LAD/UA;

the logit of the claimant count proportion in the LAD/UA;

the age/sex group;

the 12 government office regions; and

the seven super-groups under the National Statistics 2001 Area Classifications for Local Authorities of Great Britain. The 2001 Area Classification is used to group together geographic areas according to key characteristics common to the population in that grouping. These groupings are called clusters, and are derived using census data [4].

Let $\boldsymbol{\theta}$ denotes the parameter of the interest. Let $f_1(\mathbf{y}|\mathbf{u})$ be the probability density function of $\mathbf{y}$ conditional on fixed $\mathbf{u}$ and $f_2(\mathbf{u})$ to be probability density function of $\mathbf{u}$. The loglikelihood function of the binomial observation vector $\mathbf{y}$ conditional on fixed $\mathbf{u}$ and the log of probability density function of $\mathbf{u}$ are then

$$l_1 = \text{Const.} + \sum_{d=1}^{D}\sum_{i=1}^{I}[y_{di}\eta_{sdi} + n_{di}\ln(1+\text{Exp}(\eta_{sdi}))]$$

$$l_2 = \text{-}(1/2)[\text{Const.} + D\ln\varphi + \varphi^{-1}\mathbf{u}'\mathbf{u}]$$

where $\mathbf{u} = [u_1, u_2, ..., u_D]'$ follows normal distribution with zero mean and variance of $\varphi\mathbf{I}_D$.

The $\boldsymbol{\beta}$ and $\mathbf{u}$ values that maximise $l = l_1 + l_2$ are called penalised likelihood estimates. See Saei and McGilchrist (1998). The function l has been called the hierarchical likelihood by Nelder and Lee (1996). The best linear unbiased-type (BLUP-type) estimate of $\boldsymbol{\theta}$ is based on penalised likelihood estimates of the $\boldsymbol{\beta}$ and $\mathbf{u}$. Given φ_0**,** the iterative procedure used to obtain the penalised likelihood (PL) estimates $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{u}}$ (hence $\tilde{\boldsymbol{\theta}}$) can be specified as follows:

(a) Set k=0 and initial values $\boldsymbol{\beta}_0$, $\mathbf{u}_0$, and φ_0

(I) Calculate $\begin{bmatrix}\boldsymbol{\beta}_{k+1}\\ \mathbf{u}_{k+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}_k\\ \mathbf{u}_k\end{bmatrix} + \mathbf{V}_k^{-1}\begin{bmatrix}\mathbf{X}_s'\\ \mathbf{Z}_s'\end{bmatrix}\partial l_1/\partial\boldsymbol{\eta}_k - \mathbf{V}_k^{-1}\begin{bmatrix}\mathbf{0}\\ \varphi_0^{-1}\mathbf{u}_k\end{bmatrix}$

where $\mathbf{V}_k = \begin{bmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{bmatrix}\left(-\partial^2 l_1/\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k'\right)[\mathbf{X} \quad \mathbf{Z}] + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \varphi_0^{-1}\mathbf{I}_D\end{bmatrix}$

and $\partial l_1/\partial\boldsymbol{\eta}_k$, $\partial^2 l_1/\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k'$ are first and second order derivatives l_1 with respect to $\boldsymbol{\eta}_k$ and evaluated at $\boldsymbol{\eta}_k$.

(II) Set k = k+1 and return to step 2 and repeat the procedure until the estimates converge.

At convergence the BLUP-type estimate of $\boldsymbol{\eta}$ is obtained by substituting the converged values $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ in the right hand side of (1).

(b) Once iterations of (a) have converged to $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$, the ML estimate of φ conditional on the PL estimates obtained in the previous is

(III) $\varphi_{(k)} = D^{-1}(\text{tr}(\mathbf{T}^*_{(k\text{-}1)}) + \tilde{\mathbf{u}}_k\tilde{\mathbf{u}}_k)$

where $\mathbf{T}^*_{(k-1)}$ is the matrix $\mathbf{T}^* = (\varphi^{-1}\mathbf{I}_D + \mathbf{Z}_s'\mathbf{B}_s\mathbf{Z}_s)^{-1}$ with $\varphi_{k\text{-}1}$ and $\mathbf{B}_{(k\text{-}1)}$ substituted

for φ and $\mathbf{B} = -\partial^2 l_1/\partial\tilde{\boldsymbol{\eta}}\partial\tilde{\boldsymbol{\eta}}'$ respectively.

The estimation process between (a) and (b) continues until convergence in both steps.

Let $\mathbf{V} = \begin{bmatrix} \mathbf{X}'\mathbf{B}\mathbf{X} & \mathbf{X}'\mathbf{B}\mathbf{Z} \\ \mathbf{Z}'\mathbf{B}\mathbf{X} & \varphi^{-1}\mathbf{I}_D + \mathbf{Z}'\mathbf{B}\mathbf{Z} \end{bmatrix}$ and $\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$ denote the partitions of the matrix $\mathbf{V}$ and its inverse corresponding to the dimensions of $\boldsymbol{\beta}$ and $\mathbf{u}$. Replacing $\mathbf{T}^*$ by $\mathbf{T}_{22}$ in III yields REML estimate $\hat{\varphi}_{REML}$ of φ. At convergence the empirical best linear unbiased-type (EBLUP-type) estimate of η_{di} is obtained by substituting the converged values $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ in the right hand side of (1). If the parameter of interest $\boldsymbol{\theta}$ is the total number of unemployment, the EBLUP-type estimate of $\boldsymbol{\theta}$ is then

$$(2) \qquad \hat{\theta}_d = \sum_{age-sex\ classes\ i} \left\{ (N_{di} - n_{di})\hat{p}_{di} + y_{di} \right\}$$

where y_{di} is the sample value (LFS value), N_{di} and n_{di} are population and sample sizes at age-sex category i in LAD d. The mean squared error estimate is based on a first order Taylor expansion.

3 Results

We used the estimated values of β and u to obtain an estimated value for unemployment levels and rates at LAD level. The LFS estimates are design-based and they are obtained from sample values directly related to the particular LAD. Since LFS estimate of a particular LAD is based on values of variable of interest only from that LAD, it is called direct estimate in survey literature. Similarly, the estimation method of this paper (model-based) is called indirect method, since an estimate for a specific LAD is based not only on the sample values of that LAD but also values from other LADs. The estimates are calibrated to ensure consistency with the published LFS national totals for age and sex and for region and socio-economic cluster. The unemployment rates are obtained by dividing the model-based estimate of level by an estimate of the economically active population. The number of unemployed in a particular LAD is estimated by

$$(3) \qquad \hat{\theta}_d = \sum_{age-sex\ classes\ i} \alpha_{di} \left\{ (N_{di} - n_{di})\hat{p}_{di} + y_{di} \right\}$$

N_{di} is an estimate of the population for age-sex group i in area d;

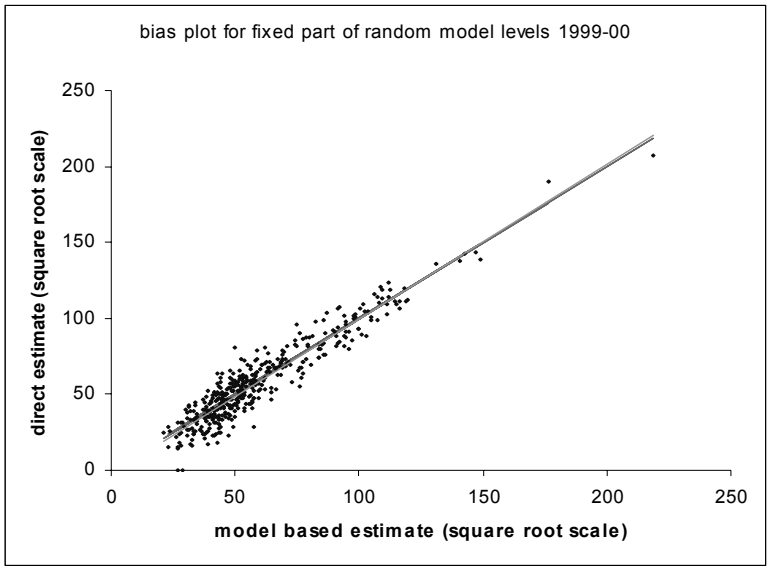
n_{di} is the number of individuals sampled in age-sex group i in area d; and

y_{di} is the number of unemployed in the sample who are in age-sex group i in area d.

α_{di} is a calibration factor.

Figure 1 is a scatter plot of the direct estimates plotted against model based estimates. As a simple bias check for the model-based estimates, a regression line was fitted. This was found not to be significantly different from the line Y=X at a 95% level. These two lines are seen to almost overlay in the figure.

Figure 1. Direct estimates regressed onto the model-based estimates.



The rate of unemployment in a LAD is simply the ratio of the number of people who are unemployed in the LAD to the number of economically active people E within that LAD. By economically active we mean people who are either employed or unemployed within the definition of the International Labour Organisation (ILO). For the purposes of the model we use the sum of the direct estimate of employed and the model-based estimate of unemployed as our estimate of E_d

Estimate of unemployment rate in LAD *d* is: - $\hat{\rho}_d = \frac{\hat{y}_d}{\hat{E}_d}$

Figure 2. Model based standard errors versus LFS direct standard errors.

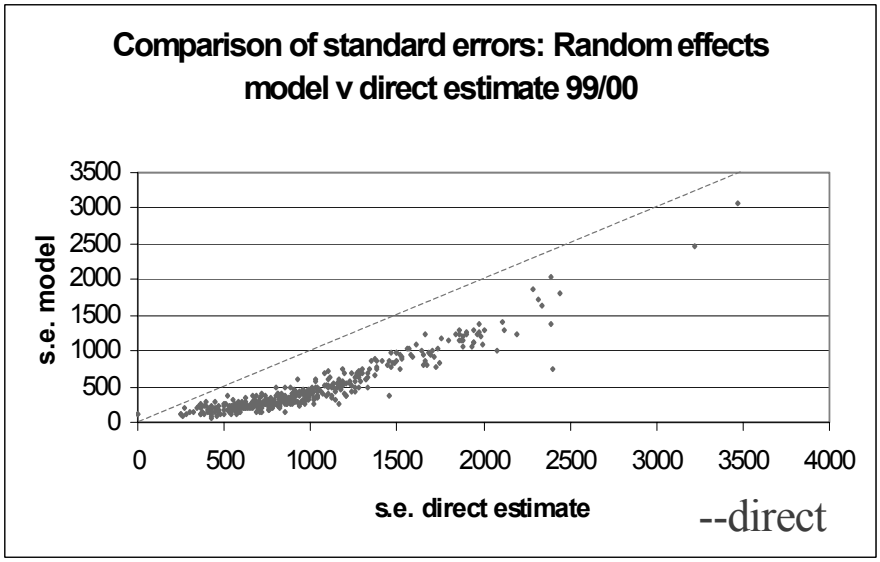
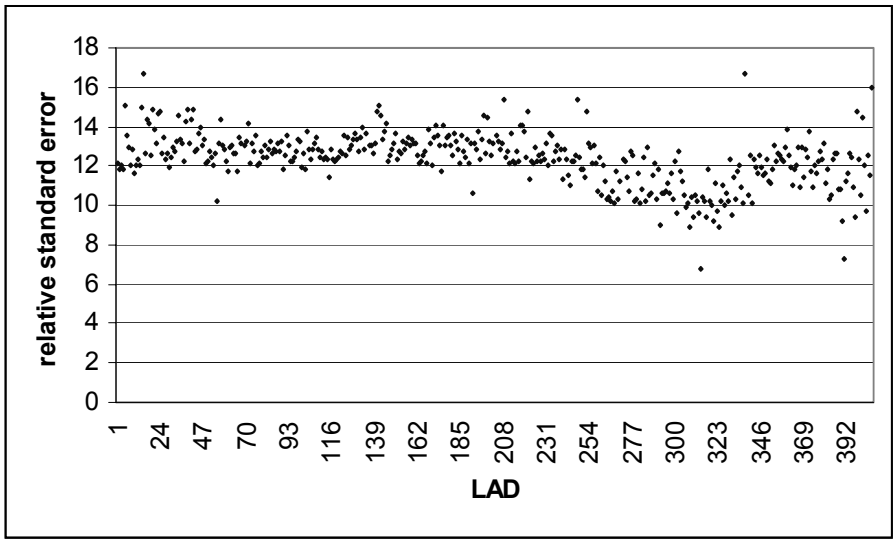


Figure 2 shows the scatter graph of standard errors of the model-based unemployment estimates against the direct standard error. The graph shows a clear advantage of the model-based estimate over direct estimate for all LADs.

The publication threshold is another parameter of interest in estimating unemployment at LAD level. The target criterion for publication of the estimates is that the relative standard error is less than 20%.

The relative standard error is defined as the standard error divided by estimate. Figure 3 below shows that this requirement has been met for all of the 406 LADs.

Figure 3. Relative standard error plotted against LAD id.



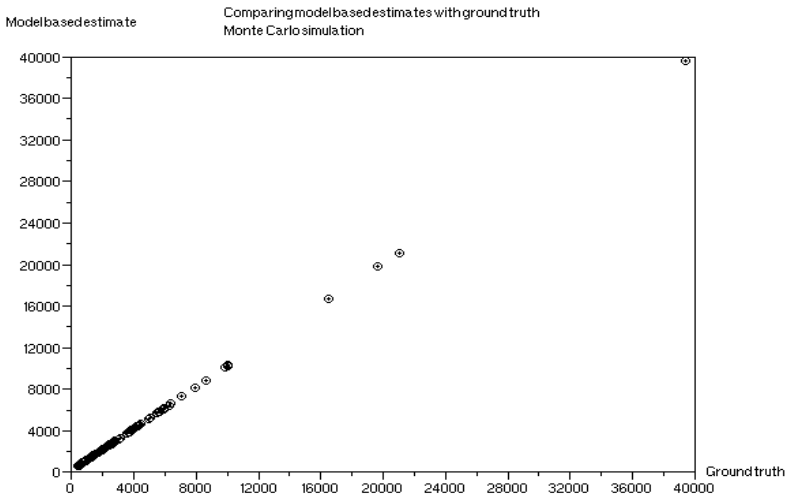
4 Simulation Study

As a further validation of the modelling methodology a simulation study was performed at ONS. In order to provide a realistic simulation, we used the LFS data as a base. Simulation including all 406 LAD and full covariate are computationally very intensive and so a representative subset of LADs and auxiliary information was selected. Data from the year 0001 were used for this simulation. A hundred LADs from a possible 406 were chosen to be included in the simulation. The covariates included in the simulation were age and sex and claimant count.

We fitted model (1) to the reduced data set. The estimated values of $\underline{\beta}$ and the variance φ were used as true values in the simulation. We generated 500 sets of area random effects values $\mathbf{u}$ from a Normal distribution with zero mean and variance of estimated value φ. We used these random effect values along with the regression coefficients to generate the numbers of unemployed in the sample and similarly the number of unemployed who were in the remainder of the population for each age-sex group within each LAD. We kept the same sample size as used in the Labour Force Survey and hence generated 500 alternative samples and unemployed populations. The total population values are sum of the sample and non-sample values.

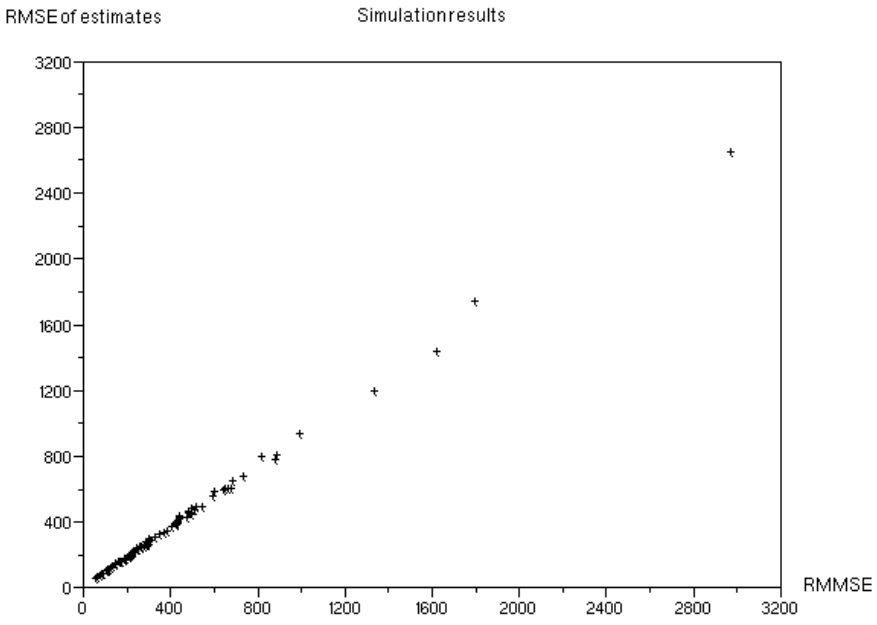
The sample values were used to obtain estimates of parameters including regression coefficients, random effect and variance of the random effect. At each simulation the number of unemployed and its associated standard error were calculated. The process of model fitting and estimation of unemployment was repeated 500 times. The average value of the estimated number of unemployed over 500 simulations was calculated for each LAD. We also obtained the average of the true values of the unemployment over the simulations. Figure 4 shows the average estimated number of unemployed against mean true value. The figure shows a close agreement between the model- based estimated value and the true value.

Figure 4 Mean model-based estimates of unemployment level against the mean true values.



The mean squared error (MSE) of the estimator was calculated at each simulation and averaged for each LAD. The square root of the averaged MSE was compared to the true root mean squared error, figure 5. The figure shows the methodology provides an estimated mean squared error that is very close to the true mean squared error.

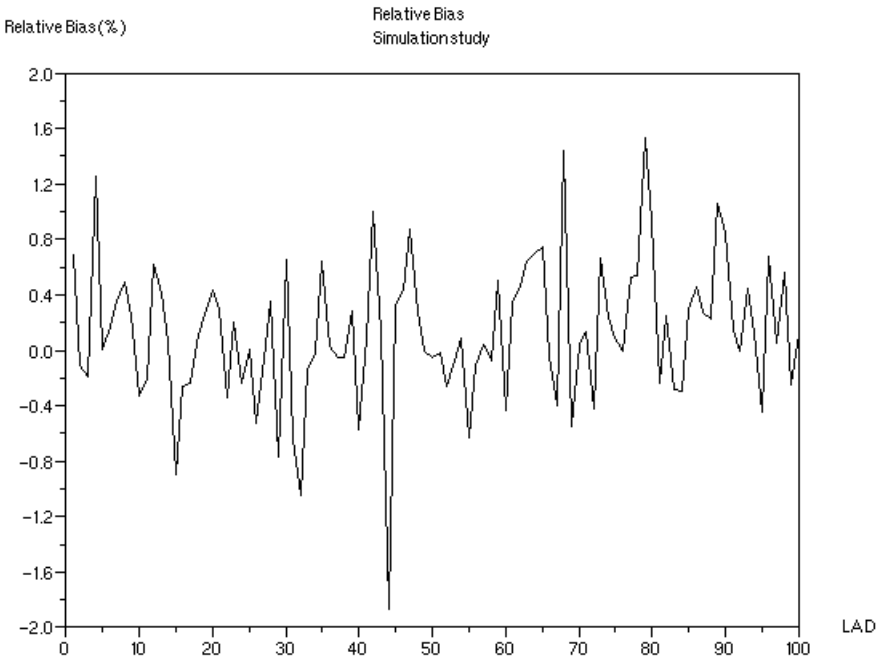
Figure 5. True root mean squared error against average RMSE.



The results of other simulations (not reported here) show that the difference between the estimated mean squared error and the true mean squared error decreases with increasing sample size.

We also calculated 95% coverage of the LAD level estimates. The results indicate that the overall average coincides with the 95% nominal level. It was also shown that the bias of the estimates were not statistically significant. Bias was generally within 1% of the estimate for any particular LAD. Figure 6 shows the relative bias of the estimates.

Figure 6. Relative bias of estimates.



The average estimate of regression coefficients and variance of random effect are very close to the corresponding true values. A summary of the simulation results is presented in Table 1

5 Discussion

The simulation has confirmed that the methodology produces standard errors of the estimates that are in close agreement with the true standard error. The estimates of the regression coefficients were shown to be unbiased. The performance of the methodology becomes even better for increased sample sizes. The random effects were assumed to be independent in both application and simulation, an extension of the model in the paper could be a model where the random effects are spatially correlated. There are many applications in which the data are available at several points in time. The methodology can be expanded to take account of the correlation between observations within a small area. The computations were carried out in SAS and codes are written by using PROC IML.

Table 1. True and REML estimate and associated standard error over simulation

Description	True value	Average	se
Variance of u (φ)	0.01272	0.0142	0.0075
Intercept	-1.298	-1.262	0.521
Age-sex factor	-2.442	-2.508	1.376
	0.972	0.925	0.615
	0.464	0.409	0.619
	2.101	2.170	1.421
	1.938	2.020	1.450

Table 1. True and REML estimate and associated standard error over simulation			
Claimant Count (CC)	0.341	0.355	0.191
CC by age-sex 1	- 0.334	- 0.348	0.264
CC by age-sex 2	- 0.078	- 0.088	0.148
CC by age-sex 3	0.077	0.064	0.151
CC by age-sex 4	0.310	0.323	0.294
CC by age-sex 5	0.158	0.174	0.294
CC by LAD	0.286	0.280	0.134

References

[1] **Brown, G., Cruddas, M. and Hastings, D.** (2003)
Development of improved estimation methods for local area unemployment levels and rates. Labour Market Trends, vol. 111, no 1, at www.statistics.gov.uk/cci/article.asp?id=372

[2] Summary publication accompanying the publication of the 2003 estimates November 2004, at http://www.statistics.gov.uk/downloads/theme_labour/ALALFS/AnnexA.pdf

[3] Labour Force Survey User Guide – Volume 6: Local Area Data. November 2004. http://www.statistics.gov.uk/downloads/theme_labour/vol6_2003.pdf

[4] **Nelder, J.A. and Lee, Y.** (1996)
Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B*, **58**, 619-678.

[5] **Saei, A. and McGilchrist, C**. (1998)
Longitudinal threshold models with random components. *The Statistician* (JRSS Series D), **47**, 365-375.

[6] **Saei, A. and Chambers, R.** (2004)
Small Area Estimation Under Linear and Generalised Linear Mixed Models With Time and Area Effects. Southampton Statistical Science Research Institute working paper MO3/15, at http://eprints.soton.ac.uk/8165/01/s3ri-workingpaper-m03-15.pdf

About the Authors

David Curtis is working in the Office for National Statistics, Segensworth Road, Titchfield, Fareham,PO15 5RR, mailto:David.Curtis@ons.gsi.gov.uk

Ayoub Saei is a senior research fellow at Southampton Statistical Sciences Research Institute, University of Southampton Highfield, Southampton SO17 1BJ, UK axs96@soton.ac.uk

Models for Data, Metadata and Knowledge

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Combining Data and Knowledge in Models. Promises and Problems

Andrew Westlake

Abstract

We collect data in order to increase our knowledge, but we always have some knowledge before we start. Our existing knowledge raises the questions for which we need more information, and it also guides us in deciding what further data to collect and how to collect it.

Models allow us to generalise from specific observed data to a wider situation. When we analyse data we (usually) update our knowledge. If we can find a formal representation for our knowledge, then a standard statistical technique provides a way to formalise the process of updating our knowledge. This can be the basis for the integration of multiple data sets that relate to different aspects of the same system.

While of general importance, this approach is the **only** way of developing a coherent and integrated understanding of complex systems which are too extensive to observe with a single data set.

But complex methodology is difficult to understand, so we must address the issues of convincing users from the application domain that our models are appropriate and valid, and of making the results obtained from the methodology accessible.

Keywords

Statistical Models, Knowledge, Uncertainty, Meta-data, Bayesian, Methodology, Opus Project, Data Integration

1 Introduction

Overview

We collect data in order to increase our knowledge, but we always have some knowledge before we start. Our existing knowledge raises the questions for which we need more information, and it also guides us in deciding what further data to collect and how to collect it. This paper explores this idea, and examines how we can formalise it. By introducing such formalisation into statistical models we can see statistical analysis and the production of statistical results as a process, rather than as a set of independent steps. We can also use this process approach to tackle the problem of coherently combining evidence from different sources which all tell us something about the same underlying system.

Knowledge and Models

Statistical ***models*** allow us to generalise from specific observed data to a wider situation. Sometimes our models are rather informal, and we think of them as assumptions (which are sometimes not made explicit). A very common assumption is that the process by which we select a sample provides a random, unbiased subset of the population about which we wish to generalise. Similarly, whenever we use a statistical method to compute a significance level or a confidence interval, we are assuming some sort of statistical distribution model. For the standard tests such as a t-test we assume Normality for the statistic. Distribution-free (or non-parametric) tests make weaker assumptions (generally based on independence), and Bootstrap methods again assume that the sample is fully representative of the population. But there always is a model, even if we do not think about it explicitly.

When we analyse data we (usually) update our ***knowledge***. This new knowledge then feeds forwards into our next data collection operation. So it can be useful to think of data collection and analysis not as a single task, but as a step in a continual ***process*** in which knowledge is continually updated as new ***evidence*** is extracted from additional data.

$$\text{Model} + \text{Knowledge}_i + \text{Evidence}_i \rightarrow \text{Model} + \text{Knowledge}_{i+1}$$

Many large-scale government business and social surveys are seen as processes, and this view is often appropriate for the analysis of more continuous data capture systems such as retail sales or traffic monitoring. This step-wise approach can also be used where we have multiple sources of evidence (i.e. multiple data sources) which are too complex to be included together in a fitting process – we just fit them in sequence. If we are concerned that the order of fitting has any influence we can iterate until the knowledge stabilises with a balance between the different sources.

How do we formalise this process approach[1] to knowledge updating? The solution is a standard part of statistical theory, which is simple to describe conceptually, if not always easy to implement in practice.

Implementation and Validation

The generic approach gives great freedom for constructing models that cover all facets of our knowledge about a system. However, this generality can cause difficulties in specification, understanding, communication and estimation. Several useful classes of model have been explored, which limit flexibility to some extent, but are easier to specify and explain. Amongst these is the class of Graphical Models (and the related group of Bayesian Networks), which are based on the concept of conditional independence. These are generally easy to conceptualise and explain, and have implementation advantages, though they do have some limitations in terms of the forms of model that are possible. As in many situations, there will be trade-offs to be made in the model specification stage between flexibility and tractability. The effect of these may need to be explored through validation.

[1] This approach applies to quantitative knowledge, such as about underlying measurements or the probability of discrete options, assessed through evidence in data. Other approaches also have their place. For example, qualitative tools such as focus groups have an important role in the generation of hypotheses, which can then be tested by quantitative means. Similarly, ontological analysis of terms and concepts is invaluable in the structuring of ideas and knowledge about classification structures. We do not discuss these other approaches further in this paper.

In any particular application domain there will be a body of generally accepted theory and knowledge about how aspects of that domain are related and interact. We can use this knowledge to build a generalised *a-priori* model of the domain that can be widely agreed and accepted. This then becomes the starting point for a more specific model of a specific system within the domain. We extend the generic model with the additional knowledge that is specific to the system to be studied.

Domain practitioners often complain that models are 'black boxes', and so are not to be trusted. This is not unreasonable, particularly with complex models of the type discussed here, and so must be addressed. We must be able to expose the structure and form of the assumptions made within a model. We must be able to provide information about which datasets were used when fitting a model, and be able to report how much and in what way they influenced the final form of the model. And we must be able to demonstrate the likely validity of the results from the model – generally, this validation will take the form of comparing the predictions from the model with actual data.

2 Models are everywhere

What sort of model?

The term ***Model*** is very widely used, and can be confusing because it implies different things to different people. Formally, a model is some abstraction (often but not always in mathematical form) representing part of the behaviour of some real-world system, selected in a particular context for a particular purpose. An often quoted remark, attributed to the statistician Prof. James Durbin, is that *all models are wrong, but some models are useful*.

Statistical Models are used to represent the relationships between observable measurements on a real system, in a way that permits estimation of (unobservable) characteristics of the real system (often referred to as parameters). A crucial component of any statistical model is the explicit choice of statistical distributions (with parameters) to represent the variability of measurements.

In computing, the term modelling is used for various processes. Data Models show the structures needed to store the various types of data used in a computer system. These are often based on the Relational or on the Object-Oriented frameworks for data structures. Process models relate to the flow of information between structures and the processes that it goes through.

Conceptual models are ways of organising the ideas (and concepts) used in some domain, together with the relationships and terminology used to refer to them. Most mental models (in our heads) are examples of conceptual models.

In many situations there are modelling frameworks that have been identified to generalise and support the process of constructing and using models. In the statistical field examples are the Generalised Linear Model framework (GLM), and the Bayesian Modelling framework. In data modelling the Relational Database Model (RDBM) plays such a role. In computing, a widely used framework is the Unified Modelling Language ([UML]). This focuses on the production of Object-Oriented computer software, but is also widely applicable for the design of structures and processes. These frameworks are all examples of meta-models, that is they are models for the process of producing models.

It is important to recognise that different models of any system can exist with different focus or with different levels of abstraction, and all can be appropriate for their intended purpose. Confusion can

arise from failing to recognise the level to which a particular construct contributes, or at which a discussion about a model is taking place.

We have found it useful to explicitly separate out those generalised models which *represent* the general knowledge about the nature of relationships and influences within a domain, in contrast to the specific and detailed models that are used to *explore* our understanding or knowledge about a specific issue or system. Thus a generalised model will make only statements about which measures are related and what the pathways of influence are, whereas a detailed model will need to be specific about the mathematics of relationship and the sources and forms of statistical variability.

Why Statistical Models?

Statistical models can allow us to generalise from specific observed data to a wider situation.

For example, in a transport system, we may count (in some reliable way) the number of vehicles using a particular segment of road, and we may stop and interview a sample of the travellers and ask what trip they are making (their origin, destination and purpose). This can tell us a great deal about what is happening at the precise location where the measurements are made on the day (or days) of observation, but, of itself, can tell us nothing about any other circumstances. If we want to generalise any of the results from the observations we need to make assumptions or otherwise formulate how our observations might relate to those on a different day, or at a different location. For example, we might assume that the breakdown of trips observed from the interviewed travellers applies to all the travellers who were not observed, and is the same on other days, even if the overall level of traffic changes. Similarly, we might establish relationships between the levels of traffic at different locations, so that measuring levels at one or more locations would allow us to produce estimates of levels at other locations.

So the purpose of creating a statistical model is to allow us to extract evidence from data about some real system in an organised and coherent way. We can then make inferences (or produce estimates) about the real system. A stochastic model will allow us to make inferences about the most likely values and variability of future observations on the system, though we will generally be more interested in estimates of underlying rates and averages. If the model is formulated in an appropriate way we can make inferences about the effect of combinations of factors for which we have no actual data.

For example, while it is at least conceptually possible to collect information (using automated systems) about the number of people entering or leaving every station on the London Underground system every day, it is impossible to conduct interviews to gather information about trip patterns at every one. However, it is perfectly possible to conduct a programme of interviewing which covers all stations over a period of time. A statistical model then allows us to bring all this information together, by making assumptions about the way in which in which demand at different points and on different days is related. We may also make use about other information, such as the connectivity between stations. The statistical nature of such a model is important, because any trip information is based on a sample of travellers, automatic counting systems may have omissions and biases, and because the behaviour of individuals is not constant from day to day.

Sometimes our models are rather informal, and we think of them as assumptions (which are sometimes not made explicit). A very common assumption is that the process by which we select a sample provides a random, unbiased subset of the population about which we wish to generalise. Similarly,

whenever we use a statistical method to compute a significance level or a confidence interval, we are assuming some sort of statistical distribution model. For the standard tests such as a t-test we assume Normality for the statistic. Distribution-free (or non-parametric) tests make weaker assumptions (generally based on independence), and Bootstrap methods again assume that the sample is fully representative of the population. But there always is a model, even if we do not think about it explicitly.

Using Statistical Models

Models vs. Data

Practitioners are often sceptical about results from models, preferring to rely on results derived directly from a particular dataset. While this attitude is understandable, it does ignore the limited applicability of a particular dataset (to what extent can the results be generalised) or the biases inherent in particular data collection methods (what do we do when different datasets give different results).

In practice, all data analysis involves some form of model, even if this is not made explicit. By making the model explicit we are better able to balance information from different sources, understand biases and generalise to the whole system.

Conflicts between datasets

The need for this comes to the fore when we have different datasets that give different answers (estimates) for the same question. For example, the UK census asked people about the location of their main work, and, from their known home location, was able to compute information about demand for travel to work between various origin and destination zones. The London Area Transport Survey (LATS) conducted household interviews at around the same time in which respondents kept a diary of their travel behaviour, from which similar origin-destination demand patterns for work were derived. The results differ substantially, often by 30%. Similarly, LATS roadside interviews ask about origin and destination of the current trip, and the demand estimates from this data are again different.

The practitioner's response to these differences is usually to ask which is right and can be trusted, and which is wrong and should be discarded.

The statistician's response is to say that (probably) they are all correct but they are just different. The source of this difference is rarely just statistical variability – in the LATS case all the sample sizes are easily big enough to be reliable. Rather the differences are due to biases (systematic differences) of some sort. Generally there are two reasons for such bias.

1. The questions are different. This is clearly the case with the census and LATS household data. The census asks about 'usual' place of work, whereas LATS records actual travel to work, which will not always be to the usual place.
2. The sample selection processes are different, so the different subgroups of the population (with different behaviour) are present in different proportions. This applies to the LATS household and roadside data, in several ways. The household survey has a rigorously defined sample selection process, but the drop-out process is biased with regard to household size. In contrast the selection process for the roadside interviews is based on randomly selecting and stopping passing vehicles, which has a different drop-out process. Similarly, whereas the household diary covers all trips, only a subset of trips can be detected at the roadside.

To obtain coherent information about a system from multiple data sources we must take into account the differences in the processes that yield the different datasets, and we do this by introducing these as factors in our statistical model.

Using results from model

How do we persuade practitioners that models (and the results from them) are valid and useful? We address this through the use of meta-data about the models and the model fitting processes. From a philosophical perspective we would argue that there is always a model, so it is better to understand it and be able to criticise it than to pretend that there is no model. However, rather than trying to win a philosophical argument we concentrate on exposing the qualities of a model so that users can make judgements as to the usefulness of model results. We focus on providing information to users about the provenance and reliability of results obtained from a model.

3 Types of knowledge

The knowledge that we have about a system can take many forms.

Knowledge about the System

When we decide to collect data about some system or process it is often because we recognise a gap or deficiency in our existing knowledge. With continuous data collection it is because the system is continually evolving and we want to update our knowledge about its current state.

When we design a survey we use our existing knowledge to decide which questions to ask, how to formulate them and how to bring them together (in a questionnaire, perhaps) in a way that is likely to yield of high quality, i.e. that accurately reflects the state of the system (or person) being questioned. If we have a choice about data collection methods then we use our knowledge to assess which will be the most appropriate (usually most cost effective) single or combination of methods to use. Finally, having identified the target population about which we want to collect information, we design a sampling process for the selection of units (usually people) about whom data will be collected. This latter is often the only factor that is explicitly associated with the collected data.

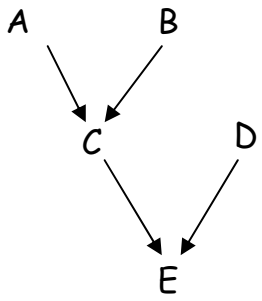
Knowledge about Values

We may know something about the range and distribution of values associated with variables: *from past data we have estimated the mean value of C to be m (with standard error s_m) and the standard deviation to be s_p. The distribution looks approximately Normal, but there is a suggestion of skewness with a long upper tail*.

We use this type of knowledge in designing data collection (for example in power calculations to determine sample size), and it can influence the choice of analysis techniques and hypotheses, but with classical approaches to analysis there is no way to use it directly within the analysis itself. Bayesian methods, on the other hand, explicitly allow such prior knowledge to be included in the analysis process.

Knowledge about Relationships

We may have ideas about the way in which factors interact and influence each other: *A and B both influence C, which in turn, with D, influences E*. We may have ideas about the form of relationship between factors: *E increases as D increases*.



We may have knowledge about the value of parameters in relationships. *We think the slope of the relationship between D and E is about 1.5, certainly near D=100* or *we think that about 30% of the variability in C can be explained by A*.

We use this knowledge to make choices about the appropriate form for the statistical analysis that we apply to the collected data. For example, using regression analysis to explore the effect of A and B on C implies the assumption that both affect C linearly and that their effects are additive.

Notice that we can include ideas about relationships (both form and context) and about distributions. The latter allows us to include uncertainty in the model. This can be related to variability in the observation process (whether inherent to measurement or from sampling) and to the detailed form of the model. We can also express uncertainty about the exact form of components of the model, generally by using wider classes of relationship or distribution and including uncertainty about the parameters that determine the specific form. These ideas are expanded later.

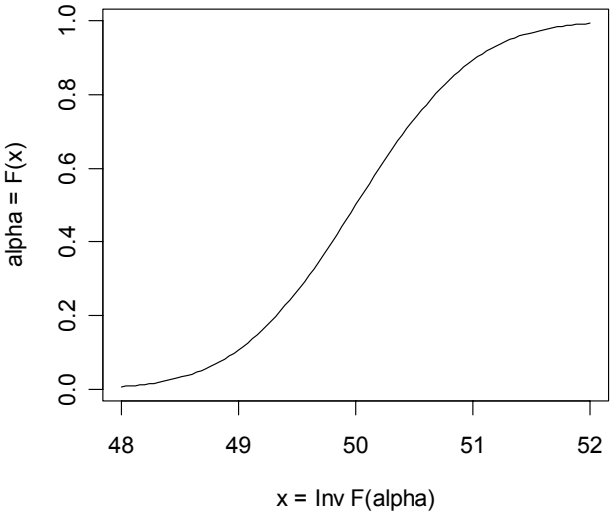
4 Representation of knowledge and uncertainty

From Confidence to Uncertainty

In simple statistical analysis we represent the uncertainty associated with an estimate of a parameter by calculating a confidence interval. For different levels of confidence we obtain different intervals (or limits) and we can represent the set all limits as a distribution over the possible parameter values. In many cases this will take the shape of a Normal distribution, because the Normal distribution is assumed for the data.

Confidence Limits

If we have an estimate $\overline{x}$ of the mean μ, then we usually calculate confidence limits for the true value of μ in the form $F_{\overline{x}}^{-1}(\alpha)$, $F_{\overline{x}}^{-1}(1-\alpha)$, where $F_{\overline{x}}^{-1}(\alpha)$ is the inverse cumulative distribution function[2] (CDF) for $\overline{x}$. This yields a confidence interval of size $1-2\alpha$. The size is the probability that the interval (which is a random variable because it



[2] Strictly speaking, the confidence limits can also be dependent on the true value of the mean, but we ignore that detail in this discussion.

is calculated from the data) actually includes the true value of μ.

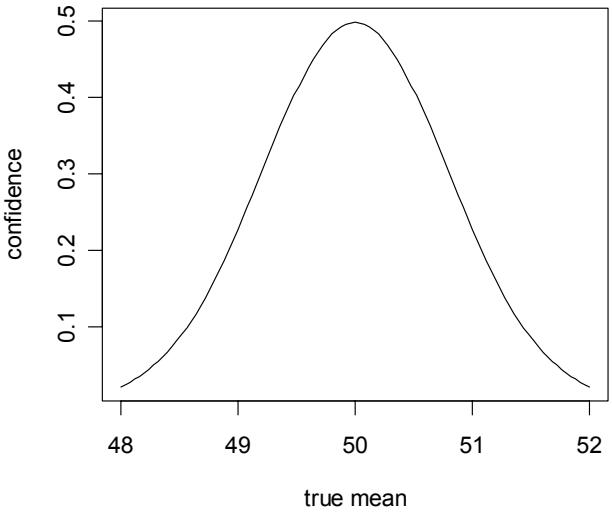
If we assume Normality for the distribution of $\bar{x}$ and flesh out the context to have n observations from which we calculate an estimate $\bar{x}$ plus a variance estimate s^2, the limits simplify to $\bar{x} \pm \frac{s}{\sqrt{n}} \Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the inverse of the cumulative distribution function of the standard Normal distribution (zero mean and unit variance).

Confidence Curves

We usually work with a single confidence interval for a given parameter. However, the size to be used for this is a matter of judgement (or just arbitrarily chosen), and in fact there is a continuous range of possible confidence intervals that could be used. An alternative approach is to represent them all.

As a concrete example, if $\bar{x}$=50, s^2=64 and n=100, the (Normal) cumulative function is as shown in the diagram on the right. The horizontal (x) axis corresponds to possible values of the true mean, and the vertical (y) axis shows the probability that the true value is lower than the corresponding x value.

This diagram can be used to obtain a confidence interval of any size – just take the appropriate probability points on the y-axis and read off the corresponding confidence limits on the x-axis. This example uses the Normal function, but any CDF can be used, as appropriate for the data.



Confidence Distribution

The same information can be shown in the form of the corresponding probability density function (PDF), because the CDF is the integral of the PDF. In this diagram, the confidence associated with any possible range of values for the true mean is the area under the graph between those values. In our example the density function is for a Normal distribution, but, again, it could be any distribution that was appropriate to the context.

The point of this is to show that the information about the true value of a parameter that we use in deriving a confidence interval can also be represented as a distribution (PDF). The cumulative form is easier to work with if we do want to compute confidence limits, but the density form can be used in other ways, as we shall see.

Uncertainty Distributions

Knowledge as Uncertainty Distribution

So far we have used the classical terminology of confidence to talk about our knowledge of the range of possible true values for the parameter. We can equally talk about our uncertainty about the true value. Then the density function above can be thought of as an uncertainty distribution – a flatter

distribution corresponds to greater uncertainty, and a more peaked one to lower uncertainty, or more precise knowledge.

Representing our uncertainty about a parameter as a distribution does not imply that the parameter is a random variable. The parameter is a fixed property of the reality about which we have collected data, and it is our uncertainty that is represented by the distribution. So we can use the mathematical properties of these functions and manipulate them as we do probabilities, but we must remember that their interpretation is not as probabilities.

We can take the idea further, and represent any uncertainty with a distribution. So far we have talked about representing the uncertainty that remains after extracting knowledge from a particular dataset by performing a standard statistical calculation. But we have knowledge (and limits on uncertainty) from other sources as well. Why not also represent this uncertainty in the form of distributions?

We do not require that the uncertainty distribution is derived from data, we can simply 'invent' it. Of course, it is not sensible to do this without some prior knowledge, or justification, to support the particular choices that we make.

For example, we could express our knowledge about a regression relationship between x and y as:

$y \sim \mathcal{N}\left(\alpha + \beta \times \left(x - 100\right), \tau\right)$, where τ is a precision[3] (or inverse variance) parameter,

$\beta \sim \mathcal{N}(1.5,1)$ – we are reasonably confident about a value near 1.5 for the slope,

$\alpha \sim \mathcal{N}(200,.001)$ – we think y is around 200 when x is 100, but are not at all confident about this,

$\tau \sim \mathcal{T}\left(.001,.001\right)$ – we know very little about the variability[4] of y around the regression line.

The first equation is a statement about the form of model and variability (Normal) for the real system, whereas the following three are statements about uncertainty over the parameters of the model.

5 Formulating Statistical Models

Model Structure

The heart of a model is a specification in mathematical terms (i.e. largely algebra) of the factors (variables) that influence the measures of interest in the system being studied, and the way in which they interact in their influence. Of course, the particular factors and form of relationships will be specific to the problem we are addressing. Our focus is on extracting coherent knowledge about the underlying system from the available data.

---

[3] This formulation (with $\tau = \frac{1}{\sigma^2}$) is often used for uncertainty, with the justification that low precision (τ near 0) is a more natural representation of great uncertainty than is a large value of the variance σ^2.

[4] The Gamma distribution (which is always positive) is often used to represent uncertainty about precision parameters.

Variables and Relationships

Variables relate to the data subjects (or various types), and their values are not generally of direct interest in themselves. We are interested in what they tell us about the underlying system (or population of data subjects).

The variables will have statistical distributions associated with them to represent variability (i.e. they are not necessarily assumed to be fixed). This variability might in part be due to measurement error, where the data recorded does not correspond exactly to the value being measured. This can happen, for example, when a vehicle counting device does not accurately register every individual vehicle, or when a respondent is uncertain about the overall income for their household. It is important not to confuse measurement error with bias in the measurement process – the latter should be included explicitly in the mathematical part of the model. Sampling issues may also be important – we return to this later.

Variability also represents unpredictability of actions and events. For example, people with the same set of background characteristics, including their jobs and home location, will make different decisions about their transport needs, in ways that are not predictable without very much greater depth of understanding (and modelling) than is feasible. And the same person may make different decisions about travel on different occasions. Also, the time taken for a particular journey may have unpredictable variations as a result of bad parking or minor traffic accidents or road works. There will also be larger effects on the travel time, consequential on the weather or the traffic loading, which we may choose to include explicitly in the model, but effects below the level of interest for the model will be treated as unpredictable variability.

We will need to be explicit about the location and nature of variability in the model, including assigning specific distributional forms (often we will assume the Normal distribution). The parameters of these distributions are included with the measures that we are interested in estimating.

Relationships can be derivations, showing how one variable is derived from others, or they can be stochastic, saying that the parameters of the distribution of a variable depend on functions of other variables (and parameters).

It is sometimes appropriate to interpret relationships as constraints. For example, a derivation equation can be thought of as an equality constraint, and a distribution where the mean is a function of other variables (as in regression) is a distributional constraint, giving the likelihood of a particular range of values.

Parameters

Conceptually, parameters relate to the true characteristics of the underlying system, as viewed through the model – they are the things about which we are trying to extract and update our knowledge. For example, in the transport context we will have parameters that relate to the probability (or rate) that people with particular demographic characteristics, living in a particular area, will want to make a particular journey for a particular purpose.

We use parameters in relationships between variables, and in the distributions we associate with variables. Familiar examples are the slope and intercept in a regression model, and the mean and variance of a Normal distribution.

Similarly, we can have relationships between parameters, and these relationships can have further parameters. Also, parameters can be defined in terms of statistical distributions, which again have

parameters. Parameters used to determine other parameters are sometimes referred to as *hyper*parameters. Depending on what is appropriate for the formulation (or parameterisation) of the model, these may not correspond to measures of direct interest, but (through the relationships) they can be used to derive those of direct interest.

Generally we will have some prior knowledge or uncertainty about the parameters, which will be more or less informative depending on what experience we can bring to the context and the understanding of the model.

The distinction between variables and parameters sometimes is a little fuzzy, particularly where the real system of interest includes components and data at different levels. So while this is a useful construct for thinking about models, we do not try to impose the distinction where it is not obvious.

Model Uncertainty

As well as uncertainty about the values of parameters in the model, we may be uncertain about the appropriate form for the model. We can cope with this by introducing parameters to control the functional form of the model, in addition to those that relate directly to the underlying system.

For example, to generalise our previous regression example of uncertainty, if we believe that a regression line may not be straight, but could be a monotonic concave or convex curve, we could use the Box-Cox transformation. This replaces x with x^{λ}, where λ is a curvature parameter – $\lambda > 1$ curves upwards, so y increases faster for larger x values, while with $\lambda < 1$ the slope declines (while remaining positive). Figure 1 shows examples of this curvature, where the precise parameterisation has been chosen to retain the slope of 1.5 when $x = 100$.

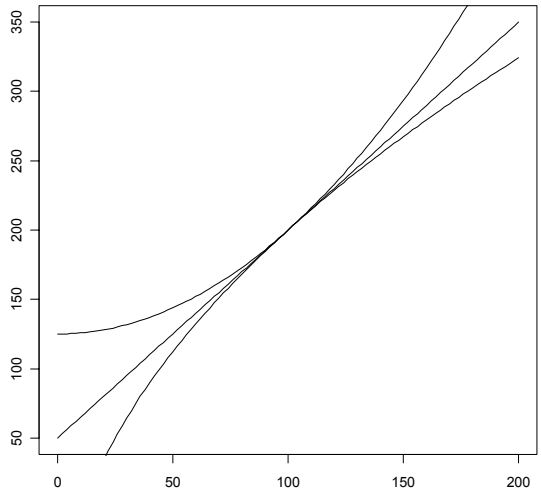


Figure 1 Box-Cox function with λ = 0.5, 1 and 2

This simple example shows how we can convert uncertainty about the form of a relationship into uncertainty about a parameter, by introducing a more general (parameterised) form of the relationship, in which our best guess is a special case. A similar approach can be used when we are uncertain about the appropriateness of particular distributional forms, since all commonly used distributions are special cases of more general forms.

Bayesian Approach

In simple statistical analysis we represent the uncertainty associated with an estimate of a parameter by calculating a confidence interval. For different levels of confidence we obtain different intervals (or limits) and we can represent the set all limits as a distribution over the possible parameter values. In many cases this will take the shape of a Normal distribution, because the Normal distribution is assumed for the data.

Although we can represent our uncertainty about a parameter as a distribution, this does not mean that the parameter is a random variable. Rather, it is a fixed property of the reality about which we have collected data, and it is our uncertainty that is represented by the distribution.

We can take the idea further, and represent any uncertainty with a distribution. Thus we do not require that the distribution is derived from data, we can simply invent it. Of course, it is not sensible to do this without some prior knowledge, or justification, to support the particular choices that we make. Where we do have knowledge about the parameter we tend to talk about knowledge rather them uncertainty distributions.

With uncertainty represented in the form of distributions, we can draw on what is known as Bayesian Methodology for working with our models.

Bayes' theorem is a simple statement about conditional probability. It comes from the recognition that a joint probability can be written as the product of conditional and marginal probabilities.

$P(A \wedge B) = P(A|B) \times P(B)$ – that is, the probability of both events A and B occurring at the same time can be calculated as the probability of B multiplied by the probability of A given that B has already occurred.

Bayes' original use of this was to show how to calculate conditional probabilities, as:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

In contrast, Bayesian Methodology uses the first formula twice to show how to reverse the ordering of the conditioning.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

If we now substitute a parameter θ for A and consider B to be our data X, we have:

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

This formula says that where we have prior knowledge about the parameter θ in the form of an uncertainty distribution $P(\theta)$, we can update this knowledge if we also know how the data distribution $P(X|\theta)$ depends on the parameter, obtaining the posterior knowledge distribution $P(\theta|X)$. Note that we do not need to evaluate $P(X)$, because it does not depend on θ, so acts as a normalising constant, and we can normalise the posterior distribution directly.

With our statistical model, if we use Bayesian methods we can say that the model gives us $P(X|\theta)$ and so tells us how to combine the evidence from the data with our prior knowledge $P(\theta)$, resulting in improved (posterior) knowledge $P(\theta|X)$. Of course, θ will be a complex structure of parameters, and we probably work in terms of density functions rather than discrete probabilities.

Updating knowledge

Given a complete model (including out prior distributional knowledge) we can predict what the results might be when we observe a particular part of the underlying system. If we have real observations on this part of the system, we can compare the characteristics of the observed data (the evidence) with these predictions. Using evidence extracted from the data we can then update (or **calibrate**) the

distributional parts of the model to bring the predictions closer to the evidence. In this way our knowledge is updated, and we can talk about the posterior knowledge after incorporating the evidence. Unlike the classical approach to model fitting which uses only evidence in the current dataset, the Bayesian approach balances the new evidence with the knowledge already in the prior distributions.

Although the Bayes updating formula can in theory be evaluated explicitly, with most complex models this is intractable. This is where the MCMC approach is used – the names Metropolis-Hastings and Gibbs Sampler are also used in this context. In effect this approach randomly generates a sequence of observations from the prior distribution of the parameter. At each step it uses the likelihood of the data under the previous and proposed parameter values in a rule that determines whether to accept the new parameter value or not. At the end of a large number of repetitions of this process we have an empirical distribution of the parameter, conditional on the data, which is an estimate of the required posterior distribution.

A number of important points should be noted

1. The model (the mathematical relationships and distributions used for calculating $\mathrm{P}\left(\mathrm{X}\middle|\theta\right)$) is not changed by the updating process, but the knowledge (which can include information about parameters that control the precise form of the relationships) is changed.
2. The updating (or fitting) process does not produce estimates from the data. Rather it updates the estimates that we can produce from the model by changing our knowledge about the parameters.
3. There is no requirement that any particular data set should contain observations on all the factors contained in the model. If a particular dataset contains no information about a parameter, then the posterior distribution of the parameter will be the same as its prior distribution, i.e. the knowledge (or uncertainty) is unchanged. We can also have factors in the model which are inherently unobservable (often called latent factors), but which influence things that can be observed.

Multiple Sources of Information

Alternative sources of information about any system often produce conflicting estimates of measures of interest. Invariably this is because the sources use different techniques to elicit and gather the information, resulting in differential biases in the data. The sources of these biases may be different survey instruments, different sample selection processes or different standards in the execution of the data collection. This presents no fundamental problem to the model-based approach.

Because all datasets relate to the same underlying system, we only need one model. But we now use multiple fitting steps to bring in the evidence from multiple datasets.

Consider first the situation where we have two independent sets of information about the same (parts of the) system, so providing information about the same (set of) parameters. Independence means that neither dataset depends on the other for a given set of parameter values, so the joint probability can be factorised into two independent parts. We can thus apply the Bayesian approach to the two datasets together.

$$\begin{aligned} P\left(\theta \mid X_1 \wedge X_2\right) &\propto P\left(X_1 \wedge X_2 \mid \theta\right) \times P\left(\theta\right) \\ &= P\left(X_1 \mid \theta\right) \times P\left(X_2 \mid \theta\right) \times P\left(\theta\right) \\ &\propto P\left(X_1 \mid \theta\right) \times P\left(\theta \mid X_2\right) \\ &\propto P\left(X_2 \mid \theta\right) \times P\left(\theta \mid X_1\right) \end{aligned}$$

This says that the posterior knowledge extracted from both sets of data can be obtained by using two fitting steps. In the first step we use one dataset to obtain a posterior distribution (from the prior knowledge), and then we use this as the prior knowledge for a second step using the other dataset. The datasets can be used in either order, and the results should be the same (to within the precision of the fitting process). This idea generalises to more than two data sets.

The assumption of independence of the datasets is reasonable for independent samples or surveys that are about the same aspects of the real system. Where there are differences in the sampling or data collection methods, for example, between household and roadside interviews, the model (i.e. the $\mathrm{P}\left(\mathrm{X}\middle|\theta\right)$) must accommodate this.

The assumption of independence may not be reasonable for all situations, and then the order of the fitting steps may make a difference. In that case we need to iterate the fitting process, using the datasets again, each time starting with the knowledge already extracted, until the posterior knowledge reaches a stable balance point.

6 The Role of Meta-Data

Meta-data as Audit Trail

Over recent years the concept of meta-data (and the recognition of its importance) has become widespread in many fields. However, the general idea of meta-data has many different applications in different areas and so means different things different people. For example, the Dublin Core proposals (and extensions such as the UK government e-GMS standard) have proved important in the context of resource discovery, especially on the Internet. Related to this is the ISO 11179 standard for meta-data repositories. Similarly, the DDI (Data Documentation Initiative) Codebook standard for the description of survey datasets has achieved wide acceptance. Several examples of the application of this approach to travel survey data are discussed by Levinson and Zofka [LeZo04]. Other authors (for example, Papageorgiou and colleagues [PPTV01]) have made proposals to extend the statistical meta-data concept to give much more complete coverage of statistical data, including sample design and tabulation.

An alternative thread that has received attention in the statistical domain is that of process meta-data. This is information that describes and documents the processes through which data has passed. This can be seen as providing an ***audit trail*** so that it becomes possible to discover details about any transformations, adjustments or corrections that have been made to data before it reaches the form in which is published. This approach to statistical meta-data is discussed by Green and Kent [GrKe02] in one of the deliverables from the MetaNet project [MetaNet].

Also from that project, Froeschl and colleagues [FGdV03] make valuable contributions about the concepts underlying statistical meta-data. Amongst their insights is the useful distinction between what they call Intentional and Extensional meta-data.

Intentional meta-data documents concepts, objectives, reasons and other factors that precede or are external to statistical data. This can include things like decisions about the sample design and data collection methods, the names and coding of variables, and the people, organisations and context associated with data. It is generally textual, and, while the structure of the components will have a formal organisation, the content will be less formally controlled.

Extensional meta-data documents actions and specifications. It includes things such as sample selection rules, derivations and transformations, file locations, process and analysis specifications. It can usually be captured by software processes, and can be part of the input specifications for other processes. The content of such items will have a tight formal specification.

Meta-data in Opus

The Opus project (funded under the ERPOS component of the European 5th framework) is developing a methodology for the integration of multiple data sources about complex systems. As part of this project we are implementing a system for reporting on the qualitative aspects of the results from the statistical methodology, as an adjunct to the results (estimates) themselves. We do this through the use of meta-data about the statistical models used.

In Opus we focus on process metadata, mostly of extensional form, to support meta-data that contains the specification of the statistical model. Our objective is to keep track of the processes that are applied in developing the statistical model from which conclusions are drawn. We assume the existence of a suitable meta-data model that covers all other aspects of the data.

Details about the Meta-data system adopted by the project appear in the following section, but the main elements are as follows:

- the mathematical specification of the model that is chosen, including all its statistical components
- the model fitting processes that are applied to the model, including all the datasets that are used
- the state of knowledge about the modelled system that is extracted from the data by the fitting processes
- specifications for the results that are extracted or reported from the final model

The intention is to capture all pertinent information about the model fitting process and link this to any results produced from the model. With this information we open up the black box of the model, so that a user can explore the qualities and reasonableness of the model and the fitting processes, and can ask questions about the reliability of results obtained from the model. Because the information is formally structured, it is also possible for other software to read the specifications and use them to repeat the model fitting process (for validation of fitting algorithms), or to apply the same model to different data.

However, while the capture of this information is essential, its mere existence is not sufficient. Facilities are needed to present the information in ways that are accessible to particular groups of user, together with guidance about the types of question that should be asked about the model and the results. This is the objective of the Reliability and Provenance concepts presented later.

7 Representation of Statistical Models and Processes as Meta-Data

Structures for Meta-data

In the Opus project we use UML to hold specifications of the structures and functionality that we have designed for handling meta-data. Figure 2 (over) shows some of the high level structures that are

relevant for this paper. The full structural model contains much more detail. Documents describing the details are available to people who sign up to join the project discussion groups, and will be widely published at the end of the project.

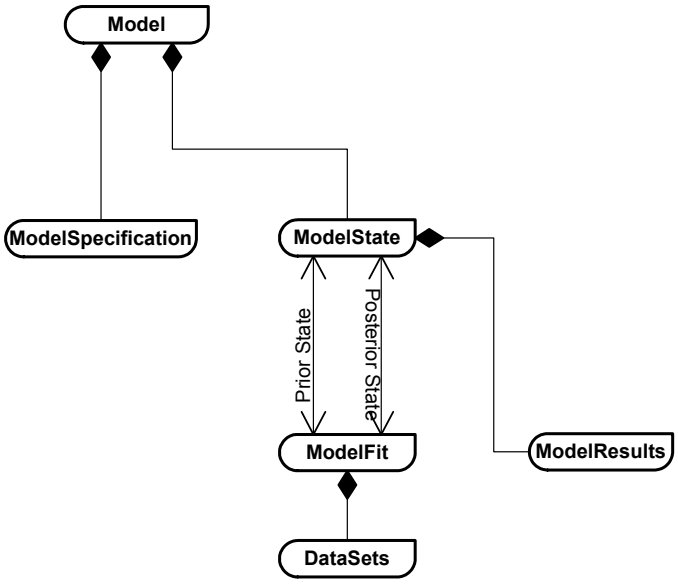
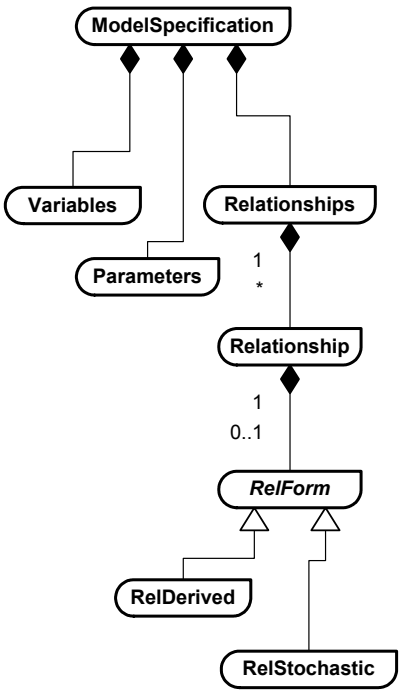


Figure 2 Outline of the Opus Meta-data Structure

The **ModelSpecification** is a (single) complex structure that contains all the information about the form of the model that has been chosen as an appropriate abstraction of the real system. This includes the variables (or factors) about the underlying real system that are pertinent to this model, the parameters that have been chosen to summarise or represent influence mechanisms in the real system, the nature and forms of mathematical and statistical relationships between the variables and the parameters, and the statistical distributions that represent the variability in observations on the system. A considerable amount of structural knowledge and expertise goes into the construction of this specification, and the specification as a whole represents the set of assumptions about the real system that are embodied in the model. The stored meta-data is mostly of extensional form, being formal specifications that can be transformed for use in suitable software, but there is also intentional meta-data that documents reasons for particular model formulations or parameterisations and for making particular assumptions.



The **ModelState** element represents knowledge about the values of parameters in the model, expressed as uncertainty distributions. Every time we use data to update (or improve) the fit of the model the knowledge changes, so in general we will have a set of states associated with the model.

The **ModelFit** represents the process of using one or more **datasets** in some well-defined methodology to update the knowledge about the system through the model. Such an updating process will start from some state of knowledge about the model (the prior state) and will produce a new state

(the posterior state) in which the knowledge (uncertainty distributions) has been updated. Often the overall process of fitting a model will involve a sequence of different fitting steps, in which different datasets are used, perhaps with different fitting methods. Iterative procedures are also possible, in which the model is repeatedly updated from various datasets until stability is reached in the uncertainty distributions. These processes produce chains of model states which represent the fitting sequence.

A fitting step may require mapping between the form of variables in the data and that in the model. For example, the model may be expressed in terms of the behaviour of individuals, but some data might only be available after aggregation. Or individual income may be represented as exact amounts in the model but only collected as banded groups in a survey. There is no problem about this, as long as it is possible to calculate the likelihood of the data that is implied by the model. In practice this means that any link between the model and data that involves variability or uncertainty needs to be represented explicitly in the model, while anything involving deterministic transformations or aggregation can be handled as a data mapping as part of the fitting step.

We assume that individual datasets are accompanied by their own meta-data describing their contents and their collection processes. In the Opus tests we will be using the DDI Codebook for this information.

ModelResults, whether conclusions, estimates or simulated data, are always based on a single state of the model, generally what might be characterised as the 'final' state after extracting all available information from all datasets. Generally speaking, results will be obtained by taking the final state of knowledge about one or more parameters and working through the mathematics of the model to be able to make statements about the implications of the model for the underlying system.

Using Meta-data with Results from a Model

Model results can always be linked back to a single state of the model, from which we have access to both the specification of the model and the chain of fitting steps that led to the final state. Thus software that is designed to support use of results from the model has access to all the meta-data that documents the final state of the model and how this was reached. This is the basis of our efforts to provide users of model results with supporting information about the provenance and reliability of the results. These efforts are discussed below.

8 Results from Statistical Models

The form of Results from Models

The end result from application of this methodology is a calibrated statistical model. This is specified in terms of a set of mathematical relationships among the variables and parameters of the model, including components that describe the stochastic variability exhibited by the underlying system. In addition, the knowledge about the model parameters that has been extracted from the evidence available in datasets is summarised in terms of posterior distributions which encapsulate the best estimates and our uncertainty about the parameters.

An experienced analyst, familiar with the methodology, can use the model to extract information about the underlying system, covering estimates of measures of interest, their variability, and the uncertainty

associated with these estimates. Where dealing directly with the mathematics of the model is too difficult, the implications of the model can be presented in the form of simulated datasets generated from the mathematical specification. A simulated dataset will generally include variability associated with the underlying system, and can also include variability arising from uncertainty about parameter values.

Provenance and Reliability of Results from Models

We anticipate the presentation of three forms of information derived from a model.

1. **Conclusions**. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.
2. **Estimates**. Presentation of the posterior distributions of quantities of interest from the underlying system. This can be done in terms of summary statistics (particularly means and standard deviations) of the posterior distributions, or of complete distributions, presented as histograms or multivariate contour plots (for example). Note that the distribution represents our uncertainty about the true value of the quantity, so it is important to present this as well as any point (best) estimates.

   Population parameters of direct interest to users (for example, in decision making) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process. But estimates can be obtained for any derivable measure on the underlying system, with a corresponding derived posterior distribution.
3. **Synthetic data**. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.

   The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations on the underlying system, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences.

These three types of information have close parallels with information obtained by more traditional methods. The difference is in the central role of the model in our methodology. Instead of presenting information that is directly derived from a dataset, and which is then inferred to be directly about the underlying system, all our information is mediated by the model. The model serves to balance and explain differences in the results obtained from separate datasets, by requiring that differences in the data collection methods or the response processes are made explicit. It also makes it possible to explore the implications of the model for combinations of circumstances for which no data has actually been observed.

For such results from a model to be useful and usable, the user must have confidence in the model. We must be able to explore and ask questions about the nature and qualities of any fitted model. We thus propose that two additional types of information should be available with all results that are derived from a statistical model.

4. **Provenance**. Information about the structure and objectives of the model (including its mathematical form), and about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.

5. **Reliability**. This relates to the posterior distributions of the model parameters. But instead of focussing on estimates of quantities of interest in the underlying system, it focuses on the uncertainty that remains about the model parameters. We explore whether the parameters are well-determined, the source of the knowledge about a parameter (ie prior knowledge or particular datasets), and how well the final model reproduces the datasets used. It is important to distinguish between *uncertainty* about parameters (which should generally decrease as more data is used or as the model formulation is improved) and *variability* in observed data that is associated with measurement processes or unpredictable behaviour.

The source of most of this information is the meta-data that describes a statistical model and that records (like an audit trail) the processes used to arrive at the final state of the model. We have proposed a structure for meta-data about statistical models that includes (potentially) all this information (it is in effect a complete audit trail for all the specifications and stages used to produce results). But we also need to find ways of presenting this additional information that are accessible and comprehensible for different groups of user.

Communicating and using knowledge

In a simple situation with a single dataset, the prior knowledge is embedded in the design of the data collection and in the head of the analyst. The evidence from the data is used to update the analyst's head, and some of this updated knowledge gets written down in reports and made available to others. This posterior knowledge may be in the form of estimates of various statistics (such as rates, means and standard deviations), or in the parameters of simple (and easily interpreted) relationships estimated from the data, such as regression slopes. What is appropriate depends on the level of skill and understanding of the analyst, and of the intended target for the knowledge.

This does not work for large and complex systems. The model is complex and covers many situations, while a dataset will refer only to a part of the system. Analysts (users of the data and knowledge) are many, with different requirements, different focuses, and different levels of understanding of the system and the model. Simply providing access to the formal specification of the model and the updated, formally expressed knowledge about the model, will not provide understanding to most domain specialists used to analysing individual datasets, so we need to find other means.

As with other complex systems (such as databases) we need to be able to provide **views** of the system (or the knowledge about it) which address the needs of specific users. To some extent we can do this by re-formulating the mathematics of the model to focus on a specific application. When doing this we can choose to leave out components of the model that are not needed in the application. This can be done by **Conditioning** (setting parameters to fixed values) or by **Marginalising** (averaging over the variability associated with the omitted components).

However, this algebraic manipulation will often not be possible with complex models, and then we have to resort to numerical methods. We can estimate averages for some output measure under differing input assumptions (ie varying input parameters) and display the results in diagrams. This is appropriate if we are wearing our analyst's hat and wish to communicate a particular message to a particular audience. More generally, we can generate synthetic datasets under various assumptions (conditioning and marginalisation, again), which we then pass over for analysis of a more traditional kind.

What can be found from such synthesised data?

Clearly, there can be no knowledge in the synthesised data that is not already in the model, because the data is generated from the model. But the domain analyst can extract knowledge from the synthesised data without needing to understand the complexity of the full model, and without having to deal with the noise in the data that would have come from the components that have been assumed out. It is clearly important to know what the assumptions have been made and to have a good general understanding of the form and limitations of the model, but this is also true (though simpler) with simple observed data sets.

In addition, since we are working from the model and are not constrained to synthesise data that corresponds to real observations, we have a number of forms of additional flexibility.

1. We can produce synthetic data that relates to combinations of factors for which we have no observations. So we could synthesise data about the flow along a traffic link at a time of year when we have no actual observations. We will have observations about the flow on the link at other times, and we have information from other places about how the flows vary over the year.
2. We can produce synthetic data for situations that do not currently exist. For example we could set factors in the model to represent the construction of a new housing development, and then produce data to investigate the impact on existing traffic flows.
3. We are not restricted to observable variables, so synthesised data can include realisations of latent (unobservable) variables.

9 Provenance and Reliability of Model Results

Objectives

Users of results from statistical models should properly be asking questions about how the results were obtained and how much confidence they should have in conclusions drawn from them. We use the term *Provenance and Reliability* to refer to this area. This covers all issues to do with the understanding and interpretation of fitted models.

Different types of user will expect answers of different complexity and detail. Some answers can be generic, describing the philosophy behind the statistical methodology and Bayesian modelling, or showing the outline of the model fitting processes (perhaps through the use of UML diagrams). Other answers will need to be based on the specific components used in the model from which the data are synthesised, and further ones will make use of the detailed posterior information about the parameters. All this information will be available in the form of metadata, the top-level structure of which has been described above. The same information may need to be presented in different ways for different types of user. Not all reasonable questions will necessarily be amenable to being answered.

Model Form

For those interested in the specification of the models, we should be able to display various components at various levels of detail. This will extend from the top level abstractions applicable to the model, right down to the details of the mathematics involved in the relationships, constraints and distributions in a particular model. Some of this should be shown in mathematical form, but graphical representations should be used wherever possible. For models that fit the Graphical Models framework, the ‘Doodle’ system in WinBugs provides a suitable style of display.

Data used

The model metadata includes links to all the data used in reaching the final calibration of the parameters, so this can be shown, and the user should be able to explore the (separate) metadata for datasets. The links between variables in datasets and those in the model are also available.

Parameters

The final model state includes information about all the posterior distributions, for the (hyper) parameters, for those induced for the parameters of direct interest and for the variables. These can be presented using standard displays of distributions, such as the graphical displays in R.

Such displays show the precision with which parameters have been determined. The reliability and suitability of the model can be explored through the progress of the parameter distributions through the calibration processes.

Domain displays

While generic displays of parameter distributions may be adequate for some statistical users, most practitioners are more used to working with specific forms of display that have been developed as particularly applicable to their domain of application. Here we face the challenge of enhancing such displays to show additional information about (particularly) reliability.

For example, in transport there are specialised displays, such as the network and Origin-Destination diagrams produced by specialised systems such as Visum. It is expected that these can be enhanced to show some aspects of variability and classification, and that they can be used to show appropriate parameter distributions, as well as distributions of actual traffic flows.

Information Requirements

Basic Areas

We have already identified the three major areas of information about a model that need to be made available to users. These are:

1. The specification of the statistical model that has been fitted.
2. The audit trail of the processes and data used to fit the model.
3. The posterior distributions of the parameters of the model, which contain all the information about the model extracted from the data.

The presentation methods specific to these areas described above will probably be sufficient for the user adept in statistical methods and mathematics. They can use these displays as tools and with them find answers to the questions that they themselves raise about the model.

Domain Users

For a user who is not familiar with statistical methodology (a non-specialist) we need to do more. We will need to provide displays that are simpler (and so do not rely on user understanding of abstract representations), and that are more focussed on the application domain of the user (so we may need different displays for different domains).

The more demanding problem, however, is that we cannot rely on the user being able to formulate appropriate questions, or even recognising that questions need to be asked. So we must address two problems.

1. How to create awareness in the user of the different nature of information obtained from a statistical model, and
2. How to provide a route map for the user through the potentially relevant questions.

For these users it is not enough to provide tools: we must provide solutions, from which they can assess the reliability of conclusions that they may want to draw from the information from the statistical model.

Creating Awareness

Within applications that we control and that provide information from statistical models, we are able to automatically introduce links and prompts to the additional information about provenance and reliability. An example of this in the context of transport networks is the Visum software.

Otherwise we rely on the original authors of the information to create awareness, rather than using software to do it automatically. Where the information from a model is used in an analysis, the analyst should already have made suitable investigations, and so can report these with the analysis and direct the reader to a suitable context for further investigation.

Where information is made available without commentary (and a major example of this is in synthetic datasets), it is necessary to make use of less direct methods, relying on the existence of metadata associated with the information. For example, this is possible for synthetic datasets placed into a Nesstar system, where the DDI metadata allows extensive commentary to be associated with the dataset. It may not be possible in other contexts. In general, we rely on the creator of the information to take every opportunity to direct users to related information about provenance and reliability.

Presentation

Once we have the attention of the user we must guide them to understanding of the nature of statistical models in general, and the reliability of specific information in a particular context.

The approach adopted in the Opus project is to develop a series of web pages that can act as a template for constructing a specific site that would support usage of a group of statistical models within some domain of application.

Some generic requirements for these pages can be identified.

1. They must provide both general guidance and specific information about the model currently of interest. So some pages will be largely static and others will be based on the model metadata.
2. Many users will be interested only in part of the model, probably related to a particular output or component of the underlying system. It should be possible to quickly focus on the relevant parameters and data (the links are in the metadata) without loosing access to appropriate guidance.
3. Particularly in models with large numbers of structured parameters (as, for example, with origin-destination pairs in a transport system) we cannot expect the non-specialist user to explore the whole range. So it is desirable that the presentation system should be able to make some automatic assessment of the reliability of different parameters (or groups of parameters), or the indication of possible anomalies. Further exploration of the Bayesian literature is needed to try to identify suitable measures for this, though we are aware of the inherent danger is such search techniques.

Evaluation of Model Reliability

Bayesian Model Checking

In all statistical modelling we face the problem of determining whether the chosen statistical model is well-suited to the reality that it represents, and whether it is well-determined by the fitting process that has been used. With classical methods we explore suitability by looking at residuals (comparing observed and fitted data values) and worrying about distributional forms (eg q-q plots), influence, etc. We address quality of fit by looking at measures such as the coefficient of determination (R^2), the residual variance and the significance levels of parameters.

Similar methods can be used in a Bayesian context, though not all have an immediate equivalent. In addition, however, we have the issue that the final posterior distributions are influenced by the initial information in the form and parameters of the prior distributions.

Gelman et al. [GCSR04] discuss model checking in a Bayesian context, and focus on the comparison of data distributions with the equivalent posterior distributions derived from the model. They point out in particular that it is not sufficient to have good correspondence in the mean and variance of a distribution – even with this it is possible to have a poor fit in the tails or inconsistent skewness – so they recommend examining whole distributions to assess model quality.

This is why we place heavy emphasis on the display of distributions as a means to understand model quality and parameter reliability

10 Conclusions

In this paper we have tried to carry through the following argument:

1. Statistical models are always needed in the analysis of statistical data, and it is better to be explicit about them than to hide them in assumptions.
2. Knowledge and uncertainty about the components or form of a model can be represented by statistical distributions.
3. Bayesian methodology enables us to extract evidence from datasets and use it to update knowledge about the parameters of a model.
4. With multiple datasets related to a modelled system, the Bayesian methodology can be applied multiple times to produce coherent knowledge about the parameters of the model.
5. Statistical models can be difficult to understand, but meta-data about the model specification and fitting processes can be used in presentations that aim to explain the provenance and reliability of results from models to users of these results.

References

[**DDI**]
Data Documentation Initiative. See www.icpsr.umich.edu/DDI for information about the DDI Alliance.

[**FWdV03**]
The Concept of Statistical Meta-Data (2003) by Froeschl, Grossmann, Del Vecchio, a deliverable from the MetaNet project.

[**GCSR04**]
Bayesian Data Analysis. Gelman, Carlin, Stern & Rubin, Chapman Hall, 2004

[**GrKe02**]
The Meta-Data Life Cycle by Ann Green and Jean-Pierre Kent. In Chapter 2 of Deliverable 4: Methodology and Tools (2002), Ed. Jean-Pierre Kent, the MetaNet project.

[**LeZo04**]
Processing, Analyzing, And Archiving Travel Survey Data, by David Levinson and Ewa Zofka. TRB 2005.

[**MetaNet**]
MetaNet: a Network of Excellence for Statistical Meta-data. See www.epros.ed.ac.uk/metanet.

[**PPTV01**]
Modeling Statistical Metadata. Haralambos Papageorgiou, Fragkiskos Pentaris, Eirini Theodorou, Maria Vardaki, Michalis Petrakos: SSDBM 2001

[**UML**]
See www.uml.org for information about UML 2.0. This is a standard developed under the auspices of the Object Management Group (www.omg.org).

About the Author

Andrew Westlake (ajw@sasc.co.uk) has a broad background in mathematics, statistics and computing, and worked for many years in university departments and in international research organisations. He is a Fellow of both the Royal Statistical Society and the British Computer Society, as well as an honorary member of ASC. He now works independently, focussing on the use of computers and databases to implement statistical processing and analysis systems.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

A Conceptual Model for Integrating Transport and Spatial Data

VS Chalasani and KW Axhausen

Abstract

Everyone involved in transport and spatial planning is at some stage involved with data production or analysis. Each transport survey is conducted for a set of objectives. Data obtained from these transport surveys do not follow any specific pattern, and are thus difficult to understand. At the same time, a research organization conducts a wide variety of surveys ranging from simple road-side interviews to the complex travel diaries, which can be either longitudinal or cross-sectional, and differences in methodology, design, and protocols will often obscure basic similarities between them. Above all, it is almost impossible to collect complete information about the existing transportation system in a single survey. Most transport surveys collect partial information and depend on other sources for more. To understand the interactions between the datasets obtained from different surveys, a conceptual data model for integrated transport and spatial data was developed. Both the transport data and spatial data were broadly classified. Transport data was classified as *travel survey data*, *transport data (infrastructure)*, and *transport data (functional)*; Spatial data were classified into *geographic data* and *geo-data.* Individual data models were developed for each classification. These data models help in streamlining the data from longitudinal surveys and standardizing the data from cross-sectional surveys. As a final step, the independent models were integrated into a single conceptual data model that represents integrated transport and spatial data. This final model facilitates easy understanding of the relationships between various data sources and allows the users to pass the information between them.

Keywords

Transport data, Geo-data, Geographic data, Transport survey, Data modelling, Entity-relationship model, Conceptual data model

1 Introduction

Improved support for the development of information systems integrating transport and geo-referenced information has been a long-term user requirement in transport planning and spatial analysis. The need to travel arises from the spatial separation of two activity locations. The continuous and mutual interaction of transport and spatial information is a central task of spatial analyses of travel patterns. Several studies have examined the spatial influences on travel patterns (Simma, 2000; Schlich, 2004), as well as the effects of various factors such as land-use (McNally and Kulkarni, 1996; Boarnet and Sarmiento, 1998), neighbourhood design (Crane and Crepeau, 1998), activity spaces

(Schoenfelder, 2003), etc. This paper focuses on the issues that require integration of transport and spatial data, and proposes a solution based on entity-relationship modelling.

Traditionally transportation professionals have collected information through various transport surveys. As technology has progressed, transport surveys have grown from simple road side interviews to complex multi-period and multi-method travel diaries. Although transport surveys are able to collect comprehensive, accurate and high quality information, they do suffer some limitations because of design and operational difficulties, as well as with respondent resistance. Transport survey data need to be enriched for the following reasons:

- Resource constraints, e.g. budget, time, etc mean that no single transport survey can collect complete information
- Outliers must be cross-checked with information from previous studies or external sources.
- Existing information, from the pre-survey process, must be integrated with the freshly observed data.

Considerable research has been conducted on transport data enrichments. Two important enrichments for the Microcensus 2000 (Swiss national travel survey: One day trip diary) were carried out at IVT, ETH Zurich are - Geo-coding the households and travel end locations (Jermann, 2003), and a study on precision of geo-coded locations (Chalasani et. al., 2004). In the first enrichment, geo-data was integrated with that of observed travel survey data to calculate the crow-fly distances. Transport network data and geographic data were integrated with the travel survey data in the second enrichment to calculate various network distances.. Spatial data was also used to augment Microcensus 2000 data at ETH Zurich with stage imputations, aggregated modes of transport, accessibility indices, travel costs, and regional traffic type (Chalasani, 2005).

Background

Recognizing the growing importance of data re-usability, ETHTDA (ETHTDA, 2005), an exclusive travel data archive, was established in 2002 at IVT, ETH Zurich. Data from several surveys ranging from simple traffic counts to the travel diaries have been archived (Chalasani, 2004). Though spatial data has been used extensively in day-to-day analyses, and in enrichments of most of the surveys, no spatial datasets were archived. Furthermore, regular updating of spatial data and data added from continuous travel surveys have increased the difficulty of understanding and mining data from divesre sources. Based on the the ETHTDA experience, and research at the institute, we have reached the following conclusions about the interaction between transport and spatial data:

- Transport surveys cannot independently collect all necessary information and must therefore be enriched
- A thorough understanding of existing transport and spatial information is a mandatory pre-requisite for any transport survey.
- High-end documentation, and dissemination of both transport and spatial data, maximises use and re-use of the data
- Linkages between transport and spatial data should be developed to support the integration of transport and spatial data.

In this study an attempt is made to develop a platform for integrating transport and spatial data. Linkages within and between transport and spatial data are explained by a conceptual data model using the entity-relationship approach. The main objective of this study is to explore the linkages between transport and spatial data through a set of interactions as relationships. This paper is organized in the following way: Chapter 2 covers data modelling issues, Transport and Spatial data are described in

Chapter 3 and 4 respectively. A conceptual data model is proposed in Chapter 5 and conclusions are stated in Chapter 6.

2 Entity-relationship modelling

The entity-relationship approach to conceptual data modeling was initially developed by Peter Chen (Chen, 1978). ER/studio 6.6.1(ER/studio, 2005) is used as a tool to draw all the entity relationship diagrams in this study. Entities are the real objects and starting point of a data model. Interactions between the entities are explained through relationships. These relationships are configured for type, existence and cardinality.

The three distinct relationship types implemented in the model are:

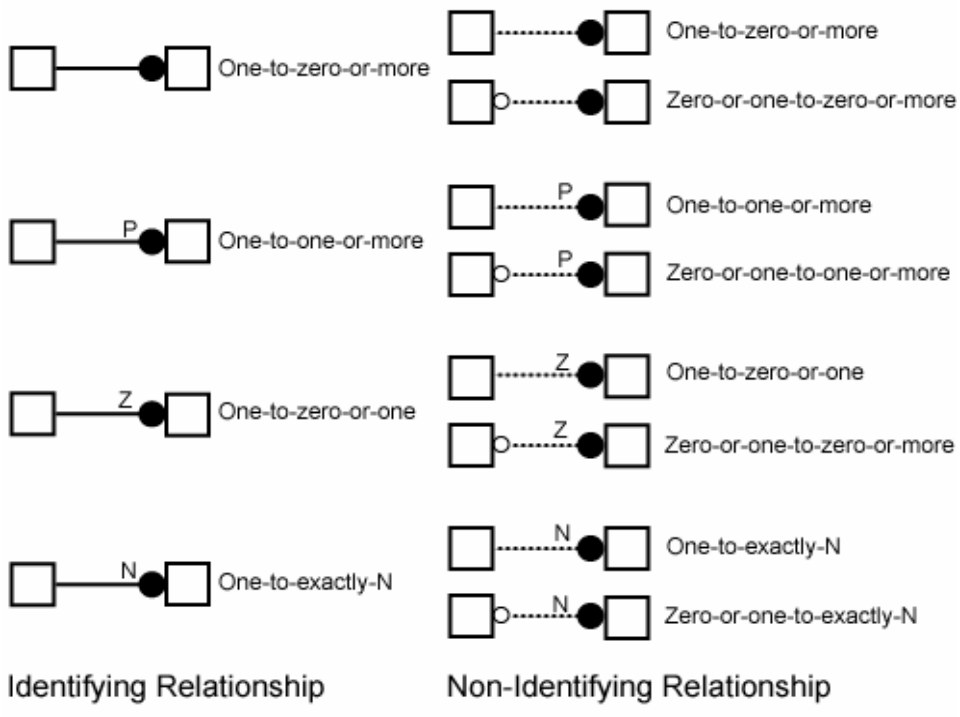
- Identifying relationships, which propagate the parent entity's primary key to the child's primary key. In the model, identifying relationships are drawn as solid lines with a solid circle terminating the child entity.
- Non-identifying relationships, which propagate the parent entity's primary key to the non-key attributes of the child. In the model, non-identifying relationships are drawn as dashed lines with a solid circle terminating the child entity. If the non-identifying relationship is optional, then a hollow diamond terminates the parent entity.
- Non-specific relationships, which denote many-to-many relationships. Because many-to-many relationships cannot be resolved, non-specific relationships do not propagate any foreign keys. In the model, non-specific relationships are drawn as solid lines with solid circles terminating both entities.

Existence describes the relationship between a pair of entities from the perspective of the child entity. Fundamentally, it asks the question, "Is a foreign key value always required in the child entity?" The possible answers are: Optional – not always, and Mandatory – always required. Existence can be enforced on the three relationship types as follows:

- Identifying Relationships, which are always mandatory.
- Non-Identifying Relationships, which can be mandatory or optional. In the model notation, optional non-identifying relationships are represented by a hollow diamond at the parent end of the relationship line.
- Non-Specific Relationships, in which existence cannot be enforced because we cannot resolve many-to-many relationships.

Cardinality describes the quantitative dimension in the relationship between a pair of entities as viewed from the perspective of the parent entity. It is read as the ratio of related parent and child entity instances. The cardinality ratio for the parent entity depends on whether the relationship is mandatory (one or more) or optional (zero or more). The model used four different cardinality ratios for the child entity: zero-or-more, one-or-more (P), zero-or-one (Z), and exactly N (N). The cardinality notation for different relationships types is illustrated in the Figure 1.

Figure 3 Notations of four cardinal ratios by relationship type



Source: ER/studio manual

Relationship existence also has implications for relationship cardinality. If a relationship is mandatory, then the cardinality must be in the form of one-to-something. If it is optional, then the cardinality would be in the form of zero or one-to-something.

Notations described in this section have been used in all the entity-relationship diagrams included in this report, but not for the spatial data hierarchy.

3 Transport data

Transport data is a generic term that covers different data types such as transportation network data, travel survey data, vehicle counts data, etc., from which comprehensive information about both the transportation system and its users can be extracted. This study classifies transport data as follows:

- Transportation system data (infrastructure)
- Transportation system data (functional)
- Transport survey data (behavioural/user reported)

The above classification is broad and general in nature, and is exclusive to this study. A detailed description of each category is covered in subsequent sections.

The following transport datasets were used in developing entity-relationship diagrams for transport data:

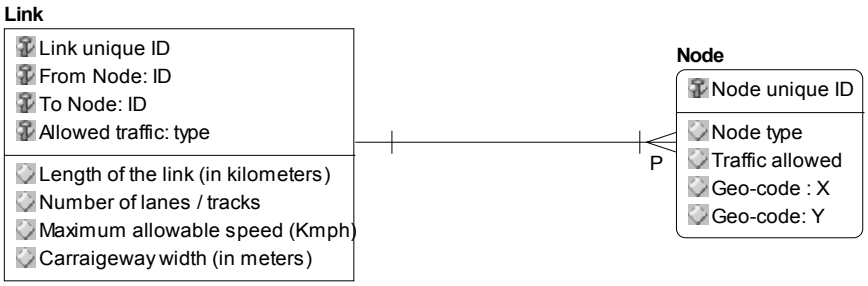
- Microcensus 2000 – One day trip dairy
- 12 weeks of leisure travel – Activity based 12 weeks diary.

- IVT national road and rail network model
- Travel module of "Household income and consumption survey" 1998
- DATELINE – long distance travel survey

Transport data (infrastructure)

Transport infrastructure data contains information about the prevailing infrastructure, i.e. the static characteristics of the transportation network, represented as a set of links and nodes, important junctions, public transport stops, etc. The transport network database consists of two data files, namely links and nodes. A simple ER diagram that represents the transport network data with two entities is shown in Figure 2. This ER diagram completely fits all the three transport network models (road network model, rail network model, and cantonal network model) available at ETH Zurich.

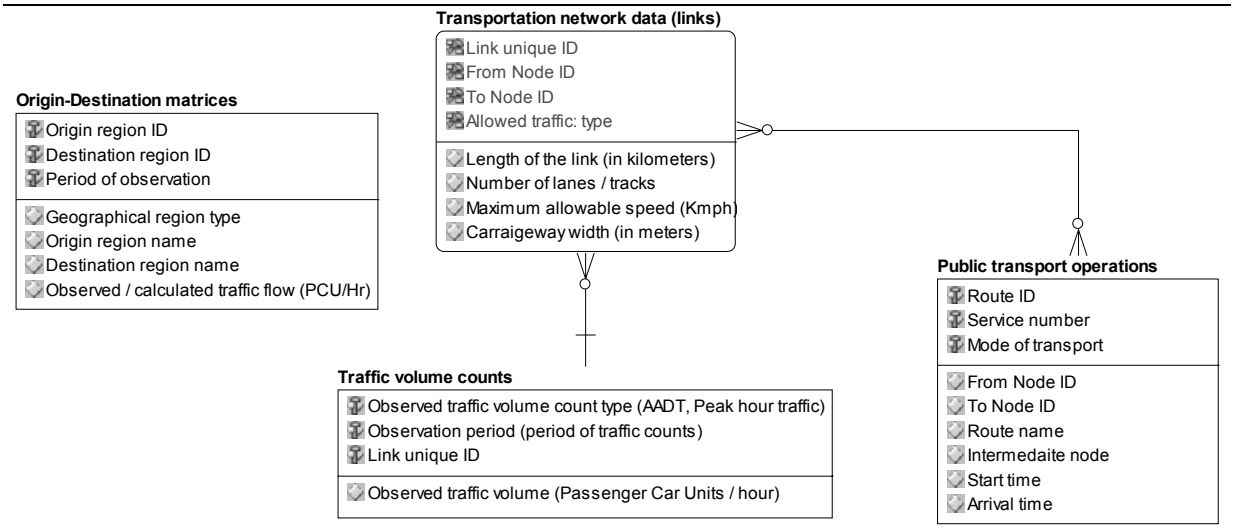
Figure 4 ER diagram for transportation network data



Transport data (functional)

Transport functional data carries information about dynamic characteristics of the prevailing transportation system. Several methods such as traffic volume counts, cordon counts, moving observer's method, etc. are used to collect the data. The functional characteristics are of two types: network operational characteristics, such as traffic movements at intersections, direction of traffic, etc., and public transport operational parameters, such as routes, schedules, frequencies, etc. A simple ER diagram for functional based transport data is shown in Figure 3. The entity "Origin-destination matrices" is not related to other entities because it is indirectly calculated from either transport survey data or traffic volume counts.

Figure 5 ER diagram for transport data (functional)

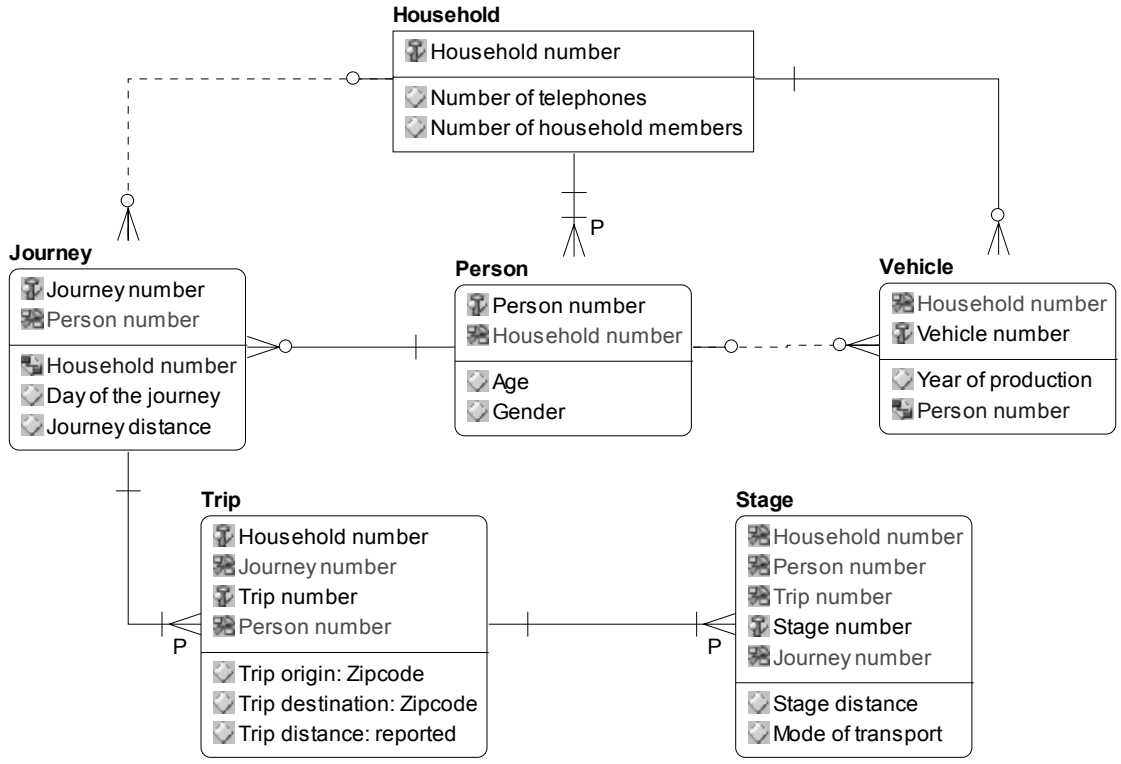


Travel survey data

This report covers both trip-based and activity-based travel survey data. After carefully editing and error checks, a travel survey data is output to set of data files. Each data file contains information on a distinct type of object, such as households, persons, vehicles, activities, journeys, trips, stages, etc.

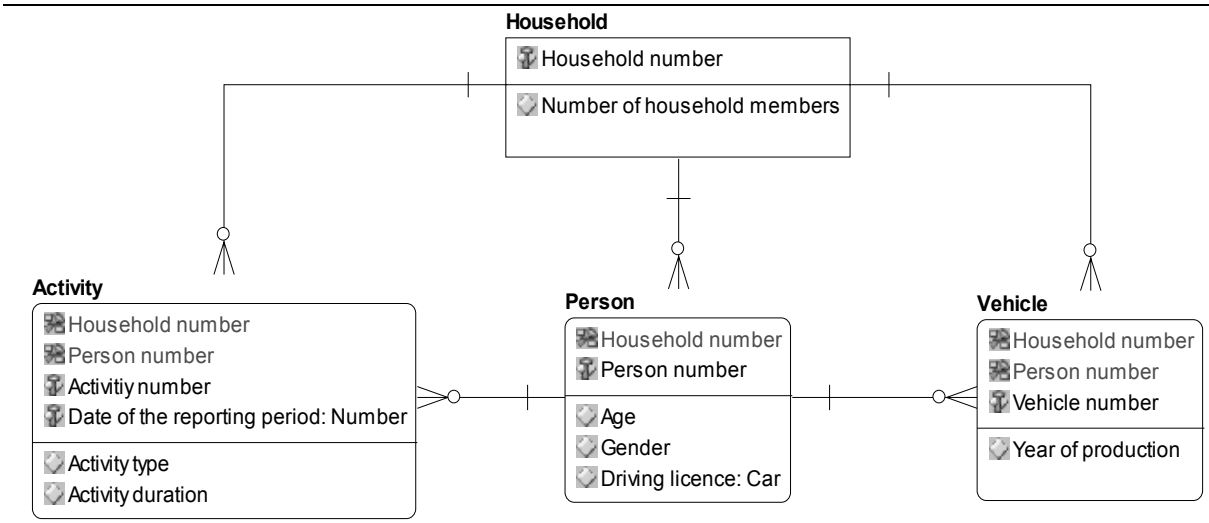
An entity-relationship diagram for a typical trip-based travel survey is shown in Figure 4. Each entity in this model is a data file. Though most of the relationships are mandatory (with solid lines), they depend on various factors such as survey context, unit of analysis, survey structure, etc. For instance, journeys can be observed at person level as well as household level, as can vehicles. As noted earlier, the structure and relationship of the ER model is survey specific. Definitions of travel terms such as ‘journey’, ‘trip’, ‘activity’ and ‘stage’, come from Axhausen (2000). This ER model represents the most used travel survey data in Switzerland, i.e. Microcensus – National household travel survey.

Figure 6 ER diagram for trip based travel survey data



Activity-based travel modelling has become more popular since the early 1990's, and is now widely used by planners all over the world. Activity-based travel survey data is much simpler in structure than trip-based travel survey data. The ER diagram shown in Figure 5 represents activity-based travel survey data from "12 Weeks of Leisure Travel", an activity survey conducted in Switzerland.

Figure 7 Entity-Relationship diagram for the activity based travel survey data



4 Spatial data

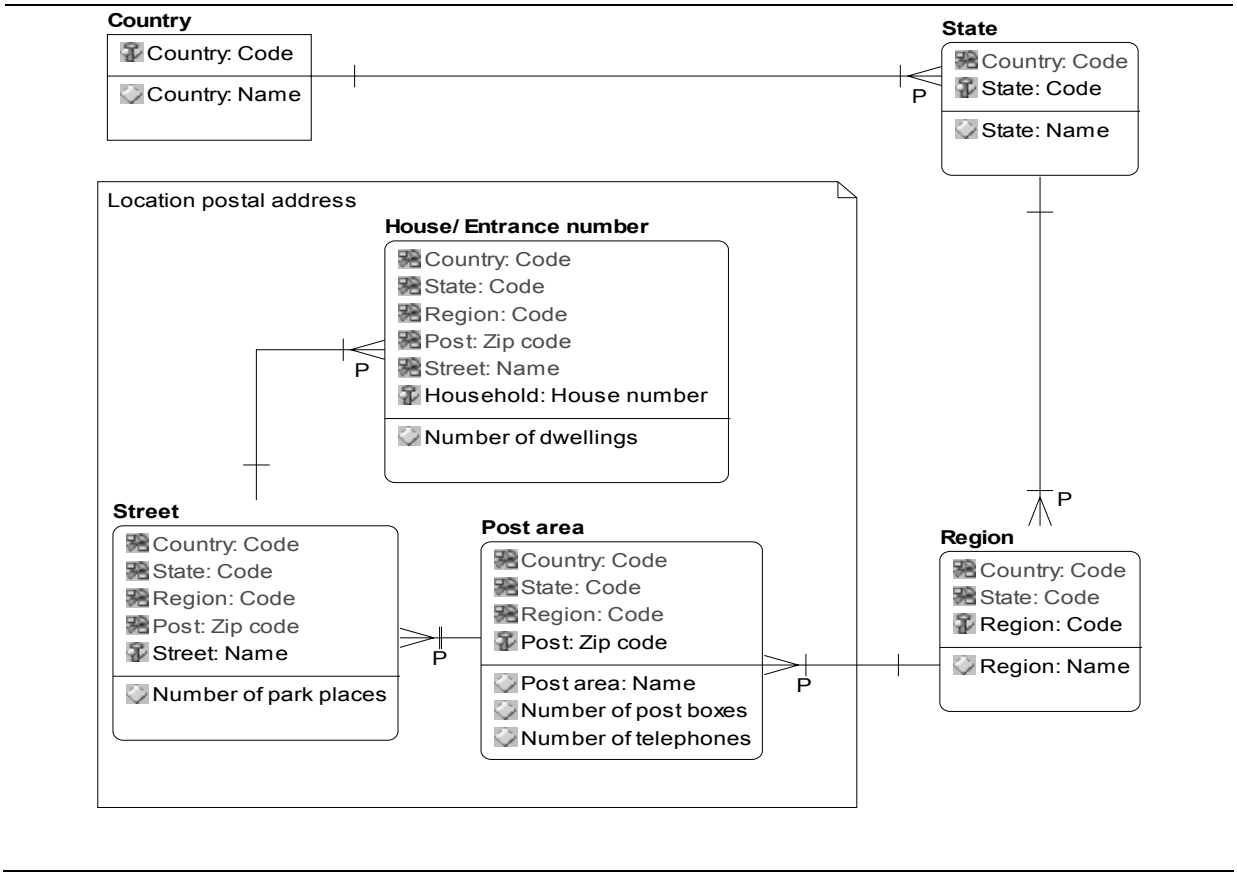
Spatial data in present context is limited to geo-referenced information i.e. geographic data and geo-data. The following spatial datasets were used in the development of entity-relationship diagrams for spatial data:

- Swiss geo-data: Geo-codes of Swiss building entrances
- Swiss geographic data (1850 – 2000): Geographic information of Switzerland

Geographic data

Geographic data is much more than electronic pictures of maps. Geographic data describes how a particular domain (continent, country, region, etct) is geographically divided according to different themes like political, administrative, transport, language, etc. Depending on its size and structure, each domain will have its own geographical hierarchy for different themes. Geographic data is less dynamic than data pertaining travel patterns. The geographical hierarchy of a country's geographic data divided for administrative purposes is shown in Figure 6.

Figure 8 ER diagram for geographic data

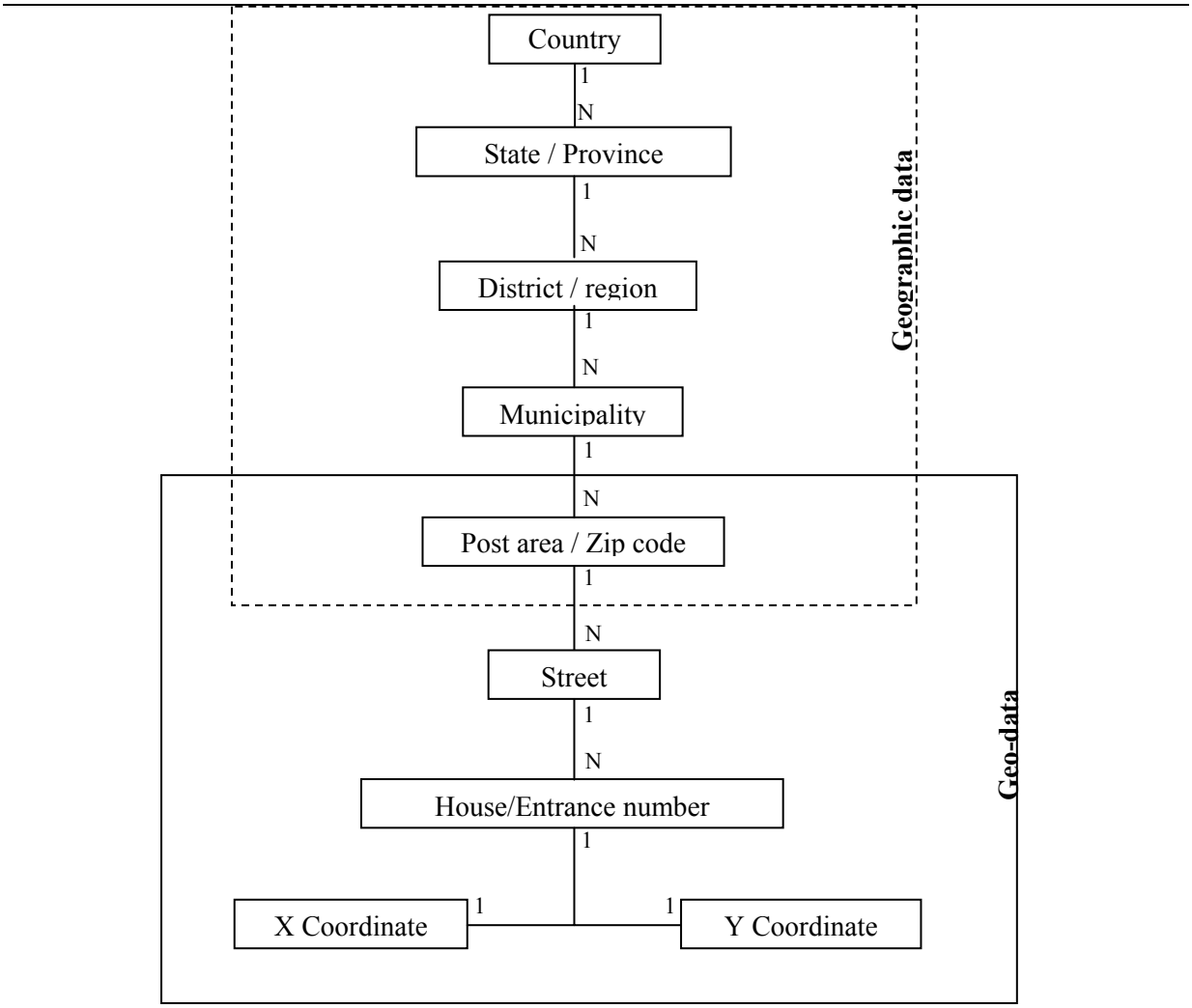


Geo-data

Geo-data is not collected by all transport surveys, but is a type of spatial data that can only be observed and collected through GPS surveys. Geo-data contains geo-information that identifies distinct physical objects such as households, building entrances, post offices, railway stations, road

junctions, etc., through a pair of geographical coordinates or geo-codes. Transport data must be enriched with geo-data to perform spatial analyses of the journeys; enrichment with geo-data will eventually be required for most the transport data. A geo-database is a database with extensions for storing, querying, and manipulating geo-data. The information hierarchy within a geo-database is similar to that of a geographic database. Figure 7 shows the spatial database hierarchy that combines geographic data and geo-data. The spatial data hierarchy differs from that of previous entity-relationship diagrams because it represents the internal structure of a single data file, while the former represents the relationships between different data files. Due to this, entity-relationship notation was note used to describe the spatial data hierarchy.

Figure 9 Spatial data hierarchy



5 A Conceptual data model for integrating transport and spatial data

The central task of this study is to develop a conceptual data model to facilitate understanding of interactions between transport and spatial data. A model using entity-relationship notation is shown in Figure 8. Each entity is a separate data file and the most important attributes are represented in the model. All the entities and relationships follow the notations explained earlier. To maintain consistency with the data classification used above, the model contains four sections: Travel survey

data, Spatial data, Transport data (functional), and Transport data (infrastructure), with a note tab used to describe the grouped entities. Descriptions of each section can be found in previous sections of this report. A logical entity "location" has been added to simplify the interactions in the model. Relationships between Households and Trips, and Households and Activities entities are optional because trips and activity data are infrequently collected for households, as compared to persons. The relationship between Geographic data and Origin-Destination matrices becomes optional when the Origin-Destination matrix's geographic region is at the lowest possible level. The relationship between Transport network data (links) and Public transport operations is optional because all links in the network need not to be accessible to all traffic types (e.g.:Public, private): on rail network links, for example, private vehicles such as cars, bikes are not allowed. Similarly, some links in a network are either for public or private transport, but not both.

Key interactions between different entities along with the key variables are listed in Table 1.

Figure 10 Entity relationship diagram for the integrated transport and spatial data with notes

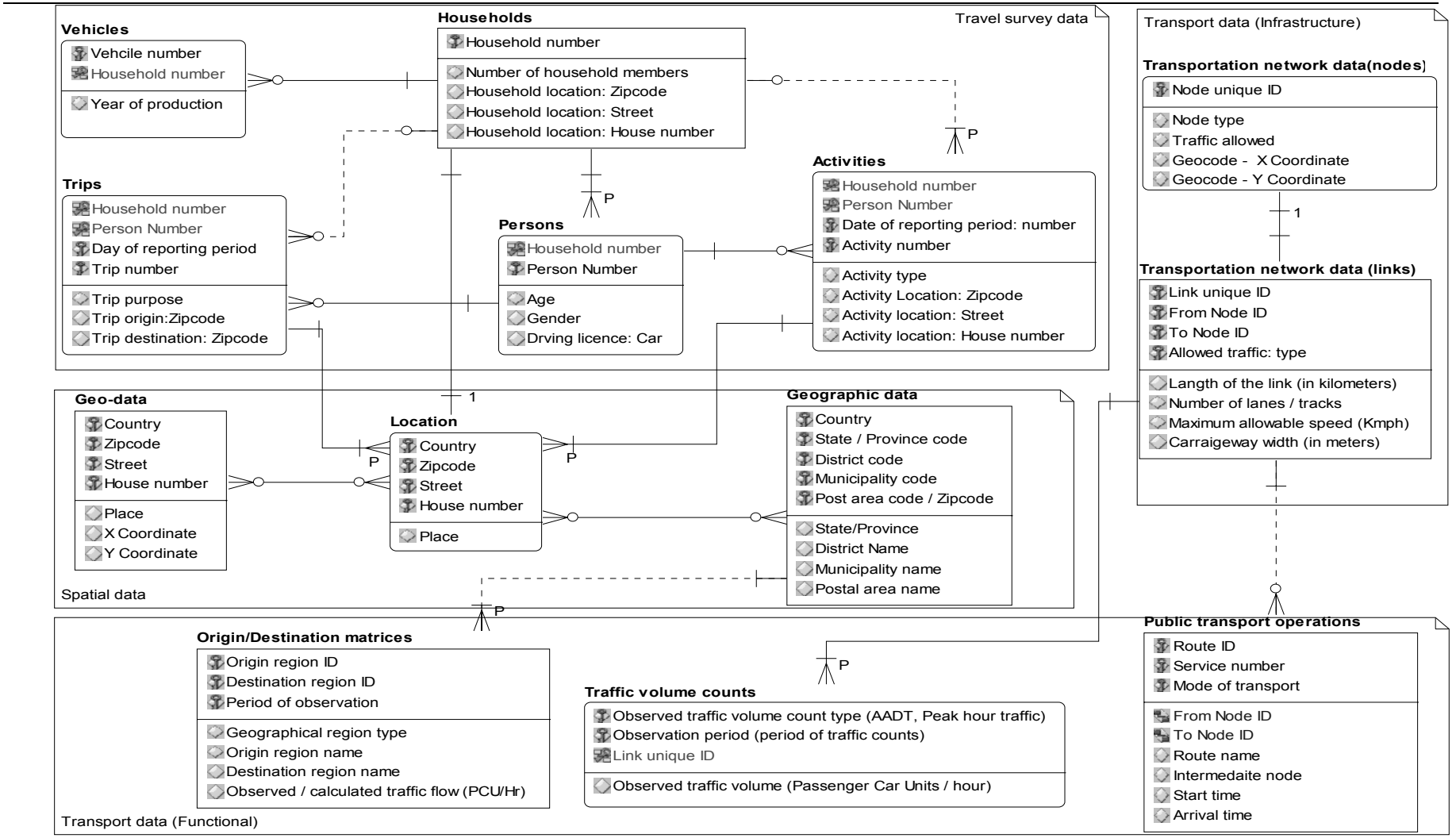


Table 1
Relationships of the conceptual data model for the integrated transport and spatial data

Parent Entity	Child Entity	Key variables	relationship
Household	Person	Household number	1 : 1 or more
Household	Vehicle	Household number	1 : zero or more
Household	Activity	Household number	1 : 1 or more
Person	Trip	Household number, Person number	1 : zero or more
Person	Activity	Household number, Activity number	1 : zero or more
Household	Geo-data	Location*	1 : 1
Household	Geographic data	Location*	1 : 1 or more
Activity	Geo-data	Location*	1 : 1
Activity	Geographic data	Location*	1 : 1 or more
Geographic data	O-D matrices	Geographic area**	1 : 1 or more
Links	Node	Node unique ID	1 : 2
Public transport operations	Links	From Node ID, To Node ID	1 : 1 or more

*: Logical entity **: Depends on the O-D matrices courseness

6 Conclusions

An understanding of the existing information (both transport and spatial) is a basic pre-requisite for any transport survey, as it not only helps in designing the information to be collected. but also improves the data quality. A thorough knowledge of existing transport and spatial data leads to better survey instrument design and improved post-processing of survey data. Basic enhancements to transport survey data are highly recommended to reduce redundancy in reported/observed transport data. When integrated with transport data, spatial data broadens the range of areas of applications, i.e. a wide range of additional problems on spatial analyses of travel patterns can be analysed. This study employed the following steps in developing a conceptual data model for integrated transport and spatial data:

- Classify the transport and spatial data for the available set of data sources.
- Develop independent data models for each sub-section of the classification.
- Identify all the possible linkages within and between transport and spatial data.
- Build a data model using the identified linkages and individual data models.

A set of datasets were used in the conceptual data model development. The conceptual data model developed in this study will facilitate:

- Integrating geographic and geo-data with both trip-based and activity-based travel survey data
- Understanding the linkages within transport data and between transport and geo-referenced data.

This model can be extended to include geographic data with other themes, e.g. such as transport regions, (transport regions), language, etc., as well as census and social network data.

References

Axhausen, K.W. (2000)
Definition of movement and activity for transport modelling, In D.Hensher and K.Button (eds.) *Handbooks in Transport: Modelling*, Transportation paper, Elsevier, Oxford.

Chalasani, V. S. (2004)
Travel data archiving: The art of presenting and preserving travel data, conference paper, 4th Swiss Transport Research Conference, Monte Verita, Ascona, March 2004.

Chalasani, V. S. (2005)
Enriching the household travel survey data: The case of Microcensus 2000, conference paper, 5th Swiss Transport Research Conference, Monte Verita, Ascona, March 2005.

Chalasani, V.S., Ø. Engebretsen, J.M. Denstadli and K.W. Axhausen (2004)
Precision of geocoded locations and network distance estimates, *Arbeitsbericht Verkehrs- und Raumplanung*, **256**, IVT, ETH Zürich, Zürich.

Chen, P.P. (1976)
The entity-relationship model – toward a unified view of data, *ACM Transactions on Database Systems*, **1**(1), 9-36.

Crane, R. and R. Crepeau (1998)
Does neighbourhood design influence travel?: bevahioural analysis of travel diary and GIS data, *Working paper*, **374**, The university of California Transportation Center, UC Berkeley.

ER/Studio 6.6.1 (2005)
http://www.embarcadero.com/products/erstudio/, Embarcadero technologies, April 2005.

ETHTDA (2005)
http://129.132.96.89/index.html, April 2005.

Jermann, J. (2003)
Geokodierung Mikrozensus 2000, *Arbeitsberichte Verkehrs- und Raumplanung*, **177**, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

McNallay, M.G. and A. Kulkarni (1996)
An assessment of the influence of the land-use transportation system on travel behaviour, *Working paper*, **96-**4, Institute of transportation studies, UC Irvine, Irvine.

Schlich, R., S. Schönfelder, S. Hanson and K.W. Axhausen (2004)
Structures of leisure travel: Temporal and spatial variability, Transport Reviews, **24** (2) 219-238

Schönfelder, S. and K.W. Axhausen (2003)
On the variability of human activity spaces, in M. Koll-Schretzenmayr, M. Keiner und G. Nussbaumer (eds.) *The Real and Virtual Worlds of Spatial Planning*, 237-262, Springer, Heidelberg.

Simma, A. and K. W. Axhausen (2000)
Mobility as a function of social and spatial factors: The case of the Upper Austria Region, Vortrag, *Land Use and Travel Behaviour*, Amsterdam, June 2000.

About the Authors

VS Chalasani (chalasani@ivt.baug.ethz.ch; tel. +41 1 633 3340) and **KW Axhausen** (Axhausen@ivt.baug.ethz.ch; tel. +41 1 633 3943) both work at Das Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, CH-8093, Switzerland.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Mind the Gap.
Metadata in e-Social Science

Phil Edwards, Karen Clarke, Judith Aldridge

Abstract

One of the problems for social science researchers trying to use multiple datasets is that concepts and classifications across these datasets differ. This is not an accident that could be prevented with more careful planning; it is in the nature of social science concepts. The challenge is to record both social 'facts' and the circumstances of their production.

In this paper, we outline a proposed solution to these problems, using a 'semantic' approach to bridging the informational gap between disparate social science datasets.

Keywords

Metadata, Semantic Web, In-Vivo Concepts, Organising Concepts

1 Towards the final turtle

It's said that Bertrand Russell once gave a public lecture on astronomy. He described how the earth orbits around the sun and how the sun, in turn, orbits around the centre of our galaxy. At the end of the lecture, a little old lady at the back of the room got up and said: "What you have told us is rubbish. The world is really a flat plate supported on the back of a giant tortoise."

Russell smiled and replied, "What is the tortoise standing on?"

"You're very clever, young man, very clever," said the old lady. "But it's turtles all the way down."

The Russell story is emblematic of the logical fallacy of *infinite regress*: proposing an explanation which is just as much in need of explanation as the original fact being explained. The solution, for philosophers (and astronomers), is to find a *foundation* on which the entire argument can be built: a body of known facts, or a set of acceptable assumptions, from which the argument can follow.

But what if infinite regress is a problem for people who want to build systems as well as arguments? What if we find we're dealing with a tower of turtles, not when we're working backwards to a foundation, but when we're working forwards to a solution?

> WSDL [Web Services Description Language] lets a provider describe a service in XML [Extensible Markup Language]. [...] to get a particular provider's WSDL document, you must know where to find them. Enter another layer in the stack, Universal Description, Discovery, and Integration (UDDI), which is meant to aggregate WSDL documents. But UDDI does nothing more than register existing capabilities [...] there is no guarantee that an entity looking for a Web

> Service will be able to specify its needs clearly enough that its inquiry will match the descriptions in the UDDI database. Even the UDDI layer does not ensure that the two parties are in sync. Shared context has to come from somewhere, it can't simply be defined into existence. [...] This attempt to define the problem at successively higher layers is doomed to fail because it's turtles all the way up: there will always be another layer above whatever can be described, a layer which contains the ambiguity of two-party communication that can never be entirely defined away. No matter how carefully a language is described, the range of askable questions and offerable answers make it impossible to create an ontology that's at once rich enough to express even a large subset of possible interests while also being restricted enough to ensure interoperability between any two arbitrary parties.(Shirky 2001)

Clay Shirky is a longstanding critic of the Semantic Web project, an initiative which aims to extend Web technology to encompass machine-readable semantic content. The ultimate goal is the codification of meaning, to the point where understanding can be automated. In commercial terms, this suggests software agents capable of conducting a transaction with all the flexibility of a human being. In terms of research, it offers the prospect of a search engine which understands the searches it is asked to run and is capable of pulling in further relevant material unprompted.

This type of development is fundamental to e-social science: a set of initiatives aiming to enable social scientists to access large and widely-distributed databases using 'grid computing' techniques.

> A Computational Grid performs the illusion of a single virtual computer, created and maintained dynamically in the absence of predetermined service agreements or centralised control. A Data Grid performs the illusion of a single virtual database. Hence, a Knowledge Grid should perform the illusion of a single virtual knowledge base to better enable computers and people to work in cooperation.(Cole et al 2003)

Is Shirky's final turtle a valid critique of the visions of the Semantic Web and the Knowledge Grid? Alternatively, is the final turtle really a Babel fish—an instantaneous universal translator—and hence (excuse the mixed metaphors) a straw person: is Shirky setting the bar impossibly high, posing goals which no 'semantic' project could ever achieve? To answer these questions, it's worth reviewing the promise of automated semantic processing, and setting this in the broader context of programming and rule-governed behaviour.

2 Words and rules

We can identify five levels of rule-governed behaviour. In **rule-driven** behaviour, firstly, 'everything that is not compulsory is forbidden': the only actions which can be taken are those dictated by a rule. In practice, this means that instructions must be framed in precise and non-contradictory terms, with thresholds and limits explicitly laid down to cover all situations which can be anticipated. This is the type of behaviour represented by conventional task-oriented computer programming.

A higher level of autonomy is given by **rule-bound** behaviour: rules must be followed, but there is some latitude in how they are applied. A set of discrete and potentially contradictory rules is applied to whatever situation is encountered. Higher-order rules or instructions are used to determine the relative priority of different rules and resolve any contradiction.

Rule-modifying behaviour builds on this level of autonomy, by making it possible to 'learn' how and when different rules should be applied. In practice, this means that priority between different rules is decided using relative weightings rather than absolute definitions, and that these weightings can be modified over time, depending on the quality of the results obtained. Neither rule-bound nor rule-modifying behaviour poses any fundamental problems in terms of automation.

Rule-discovering behaviour, in addition, allows the existing body of rules to be extended in the light of previously unknown regularities which are encountered in practice ("it turns out that many **X**s are also **Y**; when looking for **X**s, it is appropriate to extend the search to include **Y**s"). This level of autonomy—combining rule observance with reflexive feedback—is fairly difficult to envisage in the context of artificial intelligence, but not impossible.

The level of autonomy assumed by human agents, however, is still higher, consisting of **rule-interpreting behaviour**. Rule-discovery allows us to develop an internalised body of rules which corresponds ever more closely to the shape of the data surrounding us. Rule-interpreting behaviour, however, enables us to continually and provisionally reshape that body of rules, highlighting or downgrading particular rules according to the demands of different situations. This is the type of behaviour which tells us whether a ban is worth challenging, whether a sales pitch is to be taken literally, whether a supplier is worth doing business with, whether a survey's results are likely to be useful to us. This, in short, is the level of Shirky's situational "shared context"—and of the final turtle.

We believe that there is a genuine semantic gap between the visions of Semantic Web advocates and the most basic applications of rule-interpreting human intelligence. Situational information is always local, experiential and contingent; consequently, the data of the social sciences require interpretation as well as measurement. Any purely technical solution to the problem of matching one body of social data to another is liable to suppress or exclude much of the information which makes it valuable.

We cannot endorse comments from e-social science advocates such as this:

> variable A and variable B might both be tagged as indicating the *sex of the respondent* where *sex of the respondent* is a well-defined concept in a separate classification. If Grid-hosted datasets were to be tagged according to an agreed classification of social science concepts this would make the identification of comparable resources extremely easy.(Cole et al 2003)

Or this:

> work has been undertaken to assert the meaning of Web resources in a common data model (RDF) using consensually agreed ontologies expressed in a common language [...] Efforts have concentrated on the languages and software infrastructure needed for the metadata and ontologies, and *these technologies are ready to be adopted*.(Goble and de Roure 2004; emphasis added)

Statements like these suggest that semantics are being treated as a technical or administrative matter, rather than a problem in its own right; in short, that meaning is being treated as an add-on.

3 Google with Craig

To clarify these reservations, let's look at a 'semantic' success story.

> The service, called "Craigslist-GoogleMaps combo site" by its creator, Paul Rademacher, marries the innovative Google Maps interface with the classifieds of Craigslist to produce what is an amazing look into the properties available for rent or purchase in a given area. [...] This is the future….this is exactly the type of thing that the Semantic Web promised. (Porter 2005)

'This' is is an application which calculates the location of properties advertised on the 'Craigslist' site and then displays them on a map generated from Google Maps. In other words, it takes two sources of public-domain information and matches them up, automatically and reliably.

That's certainly intelligent. But it's also highly specialised, and there are reasons to be sceptical about how far this approach can be generalised. On one hand, the geographical base of the application obviates the issue of **granularity**. Granularity is the question of the 'level' at which an observation is

taken: a town, an age cohort, a household, a family, an individual? a longitudinal study, a series of observations, a single survey? These issues are less problematic in a geographical context: in geography, nobody asks what the meaning of 'is' is. A parliamentary constituency; a census enumeration district; a health authority area; the distribution area of a free newspaper; a parliamentary constituency (1832 boundaries)—these are different ways of defining space, but they are all reducible to a collection of identifiable physical locations. Matching one to another, as in the CONVERTGRID application (Cole et al 2003)—or mapping any one onto a uniform geographical representation—is a finite and rule-bound task. At this level, geography is a physical rather than a social science.

The issue of **trust** is also potentially problematic. The Craigslist element of the Rademacher application brings the social element to bear, but does so in a way which minimises the risks of error (unintentional or intentional). There is a twofold verification mechanism at work. On one hand, advertisers—particularly content-heavy advertisers, like those who use the 'classifieds' and Craigslist—are motivated to provide a (reasonably) accurate description of what they are offering, and to use terms which match the terms used by would-be buyers. On the other hand, offering living space over Craigslist is not like offering video games over eBay: Craigslist users are not likely to rely on the accuracy of listings, but will subject them to in-person verification. In many disciplines, there is no possibility of this kind of 'real-world' verification; nor is there necessarily any motivation for a writer to use researchers' vocabularies, or conform to their standards of accuracy.

In practice, the issues of granularity and trust both pose problems for social science researchers using multiple data sources, as concepts, classifications and units differ between datasets. This is not just an accident that could have been prevented with more careful planning; it is inherent in the nature of social science concepts, which are often inextricably contingent on social practice and cannot unproblematically be recorded as 'facts'. The broad range covered by a concept like 'anti-social behaviour' means that coming up with a single definition would be highly problematic—and would ultimately be counter-productive, as in practice the concept would continue to be used to cover a broad range. On the other hand, concepts such as 'anti-social behaviour' cannot simply be discarded, as they are clearly produced within real—and continuing—social practices.

The meaning of a concept like this—and consequently the meaning of a fact such as the recorded incidence of anti-social behaviour—cannot be established by rule-bound or even rule-discovering behaviour. The challenge is to record both social 'facts' and the circumstances of their production, tracing recorded data back to its underlying topic area; to the claims and interactions which produced the data; and to the associations and exclusions which were effectively written into it.

4 Even better than the real thing

As an approach to this problem, we propose a repository of content-oriented metadata on social science datasets (Aldridge, Clarke and Edwards 2005). The repository will encompass two distinct types of classification. Firstly, those used within the sources themselves; following Glaser (1978), we refer to these as 'In-Vivo Concepts'. Secondly, those brought to the data by researchers (including ourselves); we refer to these as 'Organising Concepts'. The repository will include:

- relationships between Organising Concepts - 'theft from the person' *is a type of* 'theft'
- associations between In-Vivo Concepts and data sources - *the classification of* 'Mugging' *appears in* 'British Crime Survey 2003'

- relationships between In-Vivo Concepts - 'Snatch theft' is a subtype of the classification of 'Mugging'
- relationships between Organising Concepts and In-Vivo Concepts - the classification of 'Snatch theft' corresponds to the concept of 'theft from the person'

The combination of these relationships will make it possible to represent, within a database structure, a statement such as

> Sources of information on *Theft from the person* include editions of the *British Crime Survey* between *1996* and *the present*; headings under which it is recorded in this source include *Snatch theft,* which is a subtype of *Mugging*

The structure of the proposed repository has three significant features. Firstly, while the relationships between concepts are hierarchical, they are also **multiple**. In English law, the crime of *Robbery* implies assault (if there is no physical contact, the crime is recorded as *Theft*). The In-Vivo Concept of *Robbery* would therefore correspond both to the Organising Concept of *Theft from the person* and that of *Personal violence*. Since different sources may share categories but classify them differently, multiple relationships between In-Vivo Concepts will also be supported. Secondly, relationships between concepts will be **meaningful**: it will be possible to record that two concepts are associated as synonyms or antonyms, for example, as well as recording one as a sub-type of the other. Thirdly, the repository will not be delivered as an immutable finished product, but as an open and extensible **framework**. We shall investigate ways to enable qualified users to modify both the developed hierarchy of Organising Concepts and the relationships between these and In-Vivo Concepts.

In the context of the earlier discussion of semantic processing and rule-governed behaviour, this repository will demonstrate the ubiquity of rule-interpreting behaviour in the social world by exposing and 'freezing' the data which it produces. In other words, the repository will encode shifting patterns of correspondence, equivalence, negation and exclusion, demonstrating how the apparently rule-bound process of constructing meaning is continually determined by 'shared context'.

The repository will thus expose and map the ways in which social data is structured by patterns of situational information. The extensible and modifiable structure of the repository will facilitate further work along these lines: the further development of the repository will itself be an example of rule-interpreting behaviour. The repository will not—and cannot—provide a seamless technological bridge over the semantic gap; it can and will facilitate the work of bridging the gap, but without substituting for the role of applied human intelligence.

References

Aldridge, J., K. Clarke and P. Edwards (2005)
"Bridging the gap: the case for a database of metadata", given at First International Conference on e-Social Science, June

Cole, K., K. Schürer, H. Beedham and T. Hewitt (2003)
"Grid Enabling Quantitative Social Science Datasets – A Scoping Study", online at <http://www.esrc.ac.uk/esrccontent/DownloadDocs/Colereport.pdf>

Glaser, B.G. (1978)
Theoretical sensitivity: advances in the methodology of grounded theory, Mill Valley: The Sociology Press

Goble, C and De Roure, D. (2004)
"The Semantic Grid: Myth Busting and Bridge Building", online at <http://www.semanticgrid.org/docs/ECAISemanticGrid/ECAISemanticGridFinal.pdf>

Porter, J. (2005)
"Holy Amazing Interface, Batman! Paul Rademacher's Brilliant Lodging Finder", online at <http://bokardo.com/archives/holy-amazing-interface-batman/>

Shirky, C. (2001)
"Web Services: It's so crazy, it might just not work", online at <http://webservices.xml.com/lpt/a/ws/2001/10/03/webservices.html>

About the Authors

Phil Edwards (p.j.edwards@manchester.ac.uk) is a researcher in the School of Law at the University of Manchester. His publications include contributions to *Modern Italy, Southern European Society and Politics* and *Computing*.

Karen Clarke (karen.clarke@manchester.ac.uk) is a Senior Lecturer in Social Policy in the School of Social Sciences at the University of Manchester. Her main area of research interest is the family, gender and social policy.

Judith Aldridge (judith.aldridge@manchester.ac.uk) is a Lecturer in the School of Law at the University of Manchester. Her research is primarily in the area of drug use, drug markets and gangs.The authors are joint authors of *Finding, Using and Interpreting Social Statistics: A Resource Book and Primer for Social Scientists* (Sage, in preparation)

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

European Unification through Initiative

Ken Miller, Ekkehard Mochmann, Jostein Ryssevik

Abstract

Comparative social science research in Europe is hampered by the fragmentation of the scientific evidence space. Data, information and knowledge are scattered in space and divided by language and institutional barriers. As a consequence too much of research is based on data from a single nation, carried out by a single-nation team of researchers and communicated to a single-nation audience. In order to advance interoperability, databases must be improved by metadata standards and appropriate documentation of measurement instruments. This paper will present recent developments in the field of social research, in particular the Madiera and MetaDater projects, which are laying the ground for the social science Grid and have used the Data Documentation Initiative as a building block.

Keywords

Survey metadata, DDI

1 Introduction

The Data Documentation Initiative (DDI) is an effort to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences. It benefits the research community through the rich content of the metadata that it produces, thus allowing it to be repurposed for different needs and applications. Its richness also means the metadata can be easily imported into on-line analysis systems, rendering datasets more readily usable for a wider audience.

In response to the need for better documentation, the DDI is an endeavour to provide a straightforward, consistent means for social and behavioural scientists to record clearly, and then to communicate to others, all the salient characteristics of the empirical data for which they are responsible.

This information is expressed in the most widely used specification language in the world, XML. Hence, the DDI metadata can be fully understood by computer software as well as by humans. The DDI enhances collections and aids interoperability by creating metadata that share a known structure and a specification language across many bodies of data.

The Madiera project (Multi-lingual Access to the Data Infrastructures of the European Research Area) is the latest in a series of EU projects that have utilised the interoperability aspects of the DDI standard. The project has developed an effective infrastructure for the European social science community, which improves access to and analysis of huge amounts of distributed data over the

Internet. This is achieved through a portal based on DDI metadata, Nesstar software and utilization of a multi-lingual thesaurus; Nesstar software is itself a direct outcome of an earlier EU project.

Throughout these EU projects the dependencies between the DDI and Nesstar has become increasingly evident. For a software project aiming at seamless integration between a broad range of locally controlled data holdings, the need for a generally accepted metadata standard is obvious. That metadata standards need software support to reach general acceptance might be less evident but is nonetheless true. Without software tools that can prove its usefulness and efficiency, any standard is doomed to die.

However, the DDI was initially designed as a transportation standard for data archives, basically the metadata to accompany datasets, mainly social science surveys, distributed to researchers for secondary analysis. The success of the software widened the required functionality, data coverage and also the user community. It was soon realised that the DDI needed to address the whole life cycle of social science data in all its variant forms.

The overarching objectives of the EU funded MetaDater project (Meta Data Models, Standards and Tools for Comparative Social Research) are to develop a metadata model and standards for the description of large scale comparative surveys over space and time, as well as to provide tools for metadata creation and management for such surveys. In order to achieve these objectives, the MetaDater project has developed a comprehensive life cycle data model for comparative surveys. Hence MetaDater is one of the leading projects that the DDI has turned to for input and advice.

The aim of the project is to help principal investigators, research and fieldwork institutes, large research institutes, data services and data archives to manage small to large collections of metadata and especially to make metadata capture at the source more efficient. In the long term, this should help to enhance the general quality of metadata, making comparative data over space and time much more reliable and strengthening the role of social scientists as providers of data for sound information of society.

It will bridge the gap between the minimalist documentation that is sufficient for one-time analysis of the data by the principal investigator and the extensive self-explaining documentation that is required for the optimal use of the data in further analyses. The challenge will be to overcome the heterogeneity of dataset structures and their associated documentation to allow harmonized input into present-day data browsers.

2 Background

The exchange of data between scientists is one of the major requirements of scientific progress, and critical to effective exchange is the existence of documentation that enables a full understanding of the data without any consultation with its creator. Within the social sciences it was widely thought that most documentation accompanying datasets was often inadequate.

Moreover, the unstructured and incompatible text formats of the documentation often made them technically obsolete, unable to be read by modern computer software. If the documentation of the data of the social and behavioural sciences is to be shared and reused across software, organisations, and disciplines, - that is become part of the Semantic Web -, it must use a standard specification language, and the information contained must be given well-defined and structured meaning in that language.

The aim of the DDI specification has been to address this lack of quality documentation and produce structured compatible metadata.

Whilst data itself is not a scarce resource in Europe, the availability and ability to access that data makes it appear that way. Well-developed official statistical systems combined with a variety of both academically and commercially driven data gathering programs and activities are producing a wealth of data and information about various aspects of European societies. Moreover, in the majority of European countries social science data archives have been established to secure the longer-term preservation of large parts of the available resources. These are mainly institutions that do not collect data themselves, but are there to preserve and make available for potential use what others may have collected. Still we find that availability is severely hampered by technological, judicial, economic and retrieval-related factors. Data are locked in systems, fenced by rigorous rules and treated as an economic commodity. The data are not adequately documented and usually not conceived or designed for secondary use.

If "sharing" is the most important single keyword characterising a true Grid, the key to realising the benefits of Grid computing is standardisation. Standardisation facilitates development or integration of computer software so that the diverse resources that make up a modern computing environment can be discovered, accessed, allocated, monitored, and in general managed as single virtual systems – even when provided by different vendors or operated by different organisations.

Consequently the vision of the Madiera project is to develop an effective infrastructure for the European social science community by integrating data with other tools, resources and products of the research process. This will be a fully operational Web-based infrastructure populated with a variety of data and resources from a selection of providers, a common integrated interface to the resources of existing social science data archives in Europe. Furthermore, the infrastructure will, as the Web itself, have the capacity to grow and diversify even after the initial construction period. The main objective of the project is to create a sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

However the variety of data that could be expressed with the DDI was limited, it lacked an adequate description of time series survey data and international comparative studies. This was a challenge taken up by the MetaDater project, to produce a more elaborate and extended data definition model. This work was carried out in close co-operation with and as members of the DDI Alliance.

The general management of metadata must handle comprehensive data holdings and collections of single data sets in an integrated manner. This is achieved in a relational database through normalisation, each bit of information being recorded only once, in a specific field of a specific table. That aspect distinguishes most clearly this approach from the document approach of XML files. The database has to include fields and functionalities appropriate for managing processes in which data and metadata are involved.

To organise these requirements the first product of the MetaDater project was a comprehensive data model of metadata for the outlined scope of surveys. The model covered the whole life cycle of such surveys within the logic of metadata. It gives the framework for different tasks such as managing, preserving, standardising and exporting metadata and the respective applications covering survey design and implementation phase, data collection and data processing phase, as well as including generic management and dissemination tasks with different quality protection aspects.

3 The Madiera Project

The MADIERA project (Multilingual Access to Data Infrastructures of the European Research Area) started on December 1st, 2002. It is funded by the European Commission under the Fifth Framework programme.

The overall goal of the MADIERA project is to develop an effective infrastructure for the European social science community by integrating data with other tools, resources and products of the research process.

The aim is to provide a common integrated interface to the resources of the majority of the existing 20+ social science data archives in Europe including several newly established archives in the candidate countries. Furthermore, the infrastructure will, as the Web itself, have the capacity to grow and diversify even after the initial construction period. Indeed, one of the main objectives of the project is to create a sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

The overall objectives are to provide a web based social science workbench that allows a) Data to be retrieved by browsing the catalogues of different data archives, or by performing searches across all the archives in the system; b) The metadata, documented according to the DDI standard, to be viewed with the data. c) The user to instantly run simple statistical analyses like cross/tabulation, descriptive statistics and regression; d) The data to be available for downloaded in any of a number of standard file formats; e) The display of statistics in several options, including bar-charts, time series charts, and geographic maps; f) The user to search in any language, based on thesaurus technology, and thus locate data stored in another language; g) Identification of comparative data, at both variable or datasets level, that is stored throughout Europe; h) Users to locate datasets via a map interface, using geographic coordinates in the DDI metadata; i) Identification of the same dataset stored at different locations through the establishment of a standard for naming of social science data resources; j) Access to knowledge products such as user supplied comments and links to online publications.

4 The MetaDater Project

The MetaDater project (Meta Data Models, Standards and Tools for Comparative Social Research) started on January 1st, 2003. It is funded by the European Commission under the Fifth Framework programme.

The overarching aims of the project are to develop a metadata model and standards for the description of large scale comparative surveys over space and time and provide tools for metadata creation and management for such surveys.

MetaDater will support the different stages of the comprehensive data providers and data collectors' workflow for comparative surveys to equally support extensive data documentation and publication for secondary analysis throughout the life cycle of a survey.

The main routines of the procedure to be considered in the model and tools are: a) Survey Design; b) Data acquisition, cataloguing and archiving and quality management; c) Data control and processing (including harmonisation across time and space); d) Extensive documentation on study and variable level and Metadata standardisation; e) User, data dissemination and metadata publication management.

The MetaDater project aims to produce an infrastructure instrument which facilitates metadata transfer from primary data producers to data providers and supports primary data producers and data providers in supplying the end-user efficiently and continuously with reliable, high quality, standardized and durable information about survey data. The benefits will not just be relevant for the existing institutes but also for the emerging infrastructures in Middle and Eastern European countries. It is expected that the greater integration of processes over time and various kinds of data services will bring additional incentives to preserve data and increase their usability, so the whole data and metadata production process becomes more efficient and economical. As a result, there should also be better documented data available for secondary analysis and social indicator research.

The MetaDater project will contribute to best practice in survey data resource sharing and data distribution. It facilitates next generation processing and analysis of huge amounts of data in order to increase empirical evidence and knowledge about European and global socio-economic developments.

5 Improving and Implementing DDI

The empirical observations of the social and behavioural sciences derive from surveys, censuses, administrative records, experiments, direct observation, and other systematic methodologies for generating empirical measurements. They may pertain to individual persons, households, families, business establishments, transactions, countries, and many other subjects of scientific interest. The observations may consist of measures taken at a single point in time in a single setting, such as a sample of people in one country during one week, or they may consist of repeated observations in multiple settings, including longitudinal and repeated cross-sectional data from many countries, as well as time series of aggregate data. The DDI specification has been designed to fully encompass all of these kinds of data and to provide all the information a potential data analyst needs.

However, the structure supplied by the tagged file of a DDI XML metadata record is perhaps the greatest strength of the standard, in that in allows computer manipulation of the information contained within those tags. This structure allows multiple usages of the information stored within the tags: a) the ability to input the data directly into software packages b) the tailored display of the information through style sheets to satisfy unique user needs c) the ability to perform complex precision searches d) the output of traditional style codebooks.

Basically, the DDI produces a single document with multiple purposes in which changes made to the core document will be passed along to any output generated.

Modern democracies produce a growing database for empirical social research. This data increase is only manageable with data and metadata management instruments that make the preparation of data files for access and further analysis more efficient. So far, there exists no comprehensive system that integrates the functionalities required for metadata standardisation, storage and output into the workflow.

The MetaDater project will extend the DDI standard to support the stages of data provision and collection and provide extensive data documentation throughout the life cycle of comparative surveys. From survey design and collection through to data acquisition, cataloguing, archiving and re-use, MetaDater will. extend DDI beyond a single instance and harmonise data across time and geography and will provide metadata publication management at variable, as well as study level.

One source of difficulties with an adequate documentation of a survey and data is the lack of an integrated instrument to capture related metadata when they first appear and in a standardised format. Collection of metadata starts with the survey design and the development of suitable instruments (questionnaire; collection instrument), continues with its implementation in the field and may find a preliminary end when data are validated or analysed for the first time or if new variables are constructed in this phase.

To make data documentation as efficient as possible the MetaDater project will supply applications that will support the workflow, facilitating the structured entry and storage of information when a survey is born and launched (e.g. survey design; fieldwork documentation; social and cultural survey context).

Basic management facilities are supported as well as routines to aid the editing of multilingual questionnaires with the creation of a related variable structure. It will provide the base for the data processing and documentation of cross-section datasets and comparative data. Comments on the data can be captured at any stage to facilitate the management of, or information on, problems with specified variables or codes.

The Madiera project, in contrast, has used the present DDI standard as one of the three main components on which the Madiera portal is based. The other two being a comprehensive multilingual thesaurus, ELSST (European Language Social Science Thesaurus) and the Nesstar technology for making data resources available on the web

The DDI metadata standard, supplied with a tag-library and implemented in XML, presents the structure and the possibilities. To put it into actual use Madiera needed to supply it with a lot of detailed definitions, controlled vocabularies for certain elements to add machine-functionality and Web-accessible semantics to DDI-described data. Madiera, through CESSDA (Council of European Social Science Data Archives), has been working to develop a European implementation of the DDI standard, with a common agreement on additional vocabularies, a common “template” of mandatory and recommended elements and some generally agreed upon best practice.

Language barriers are major obstacles to efficient resource location and utilisation across the European Research Area. This is especially so for comparative research that normally requires data and resources from more than one language community. Apart from a handful of significant comparative data collections that are available in several languages, the majority of sources describing European societies are only documented in one language.

If comparative data resources can be efficiently identified across language barriers, the first hurdle is already passed. This can be achieved by the use of language-independent classifications of resources as well as language-independent and thesaurus-supported application of keywords and terms to the relevant parts of the metadata records. In the practical implementation of the DDI in a multi-language Europe, the thesaurus ELSST stands out as the single most important component.

The keywords assigned to the metadata from ELSST can be instantly translated back into the supported language of the user. Initial full translation of the returned resources might then be achieved by applying standard automated Web-based translation services. We know that the quality of these translation services still do not meet scientific standards, but they might be used as a first pass in order to decide whether the use of human-powered translation might be worthwhile. And the data-location and retrieval purpose is not dependent upon the full and optimal translation service.

The ELSST thesaurus at present covers core concepts in social science research and methodology for nine European languages, English, French, Spanish, German, Greek, Norwegian, Danish, Finnish and Swedish.

The Nesstar technology has been developed through the EU-financed NESSTAR (Networked Social Science Tools And Resources) and FASTER (Flexible Access to Statistical Tables for European Research) projects. It is a state-of-the-art suit of software tools developed to run real-life data services at data archives and other large organisations.

The four basic facets of Nesstar functionality are resource location, metadata browsing, on-line analysis and data download. Within the Madiera project the goal has been to further refine the available technologies to make the software even better suited as a tool for European comparative research.

6 Conclusion

The DDI can serve as the foundation for content, distribution, use and preservation of data collections in the social and behavioural sciences, across institutions, countries, and disciplines. That foundation will be stronger if the specification is independent of any particular software or computing platform. Expressing the specification as a generalised conceptual data model will further enhance this independence. The data model is extensible and modular, supporting the specification of even the most complex data systems in a way that is simultaneously flexible and rigorous.

In further pursuit of the goal of wide adoption, the project is seeking cooperation from both data producers and statistical software manufacturers. It is hoped that DDI metadata can soon be produced by standard computer-assisted interviewing software and be accessed directly by many statistical software packages for purposes of data definition. When these two aims are achieved, the DDI specification can readily become the basis for the entire research process, from generation of a data collection instrument to production of research articles.

A further challenge will be the extension of the DDI standard to support more complex data. The current specification is excellent for documenting independent survey files, but further work is required to build on mechanisms already included to support aggregate data and hierarchical files.

The MetaDater project will facilitate metadata transfer from primary data producers to data providers and support them both in supplying the end-user with reliable, high quality, standardised and durable information about survey data. The MetaDater also supports long-term preservation of data and related metadata, to keep the "digital heritage" in the field of social research. The benefits will not just be relevant for the existing institutes but also for the emerging infrastructures in Middle and Eastern European countries.

It is expected that the greater integration of processes over time and various kinds of data services will bring additional incentives to preserve data and increase their usability, so the whole data and metadata production process becomes more efficient and economical. As a result, there should also be better documented data available for secondary analysis and social indicator research. It will contribute to best practice in survey data resource sharing and data distribution. It facilitates next generation processing and analysis of huge amounts of data in order to increase empirical evidence and knowledge about European and global socio-economic developments.

The DDI is working towards a more modular and extensible version of the standard, which will consist a conceptual model, as well as XML schemas that are derived from it. In the construction of the conceptual model the DDI has taken on board the MetaDater life cycle model for metadata standardisation. MetaDater is expected to strengthen and develop the European technological infrastructure for the social sciences and to facilitate access to well documented data for its users. It addresses the increasing request for access to comparative data across the European Union and worldwide. Currently the data model and metadata standards are being discussed with DDI expert groups and other professional organisations committed to promotion of comparative survey research such as the Comparative Survey Design and Implementation (CSDI) group.

The Madiera social science portal links together a set of separate Nesstar servers, using the common DDI standard to publish and maintain their data holdings which are then directly accessible on the Web through the HTTP protocol. The portal automatically matches resources with terms from the multilingual ELSST thesaurus. Hence European social researchers can easily locate data resources published by any of the participating data archives by browsing or searching in their preferred language. However, the multilingual thesaurus is not only a key resource for the data location process, but also for the documentation and publication process, hence Madiera has developed publishing tools where the thesaurus is built in to facilitate automatic insertion of keywords at study or variable level.

The Madiera portal has also utilised the DDI structure and content for menu driven browsing of the metadata and to offer both a geographical/map-based resource location tool and a specialised search for comparative data. The Madiera geographical interface will allow the user to circle or mark one or more areas on a map and retrieve lists of resources covering or deriving from the selected areas. In the Madiera comparative interface the end-user will be given a high degree of flexibility in defining the relevant set of comparability criteria. The system will then identify comparable data sources that potentially address the same problem areas, questions, variables, geography, sampling procedure, etc.

There is an important connection and interdependence between grid computing, data usage, standardisation, standards development and accompanying software. An important component of the Madiera project has been to develop a programme to promote the publication of DDI content. This programme has collaborated closely with the work of the CESSDA DDI Group with a mandate of developing a common template and controlled vocabularies for the DDI. However, the project is also looking outside of CESSDA to other data producers, data users and research communities to publish data, and hence contribute to the growth of the infrastructure. This extended infrastructure would hopefully then be based on a concept of metadata being seen as a collection of information that is developed and enriched through the life cycle of the dataset.

The DDI serves the social science community well with a specification that produces quality metadata with multiple purposes. It fully documents the details of datasets, it is user friendly and accessible, it integrates into the infrastructure of the Web and it supports automatic generation of statistical software system files. The widespread adoption of the DDI will vastly improve access to a range of varied datasets. Expanded use will greatly enhance comparative research; the ability to harmonise datasets over time and geography will lead to significant improvement in our understanding of societies. Increasing the availability of high-quality data is a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and that is what the Data Documentation Initiative delivers.

References

The Madiera Project Web site (http://www.madiera.net)

The MetaDater Project Web site. (http://www.metadater.org)

Alvheim, A. and Ryssevik, J.(2005)
"MADIERA, Multilingual Access to Data Infrastructures of the European Research Area" Paper presented at the First International Conference on e-Social Science, Manchester, UK, June.

Jensen, U. and Mochmann, E. (2005)
"MetaDater: Meta Data Models, Standards and Tools for Comparative Social Research" Paper presented at the First International Conference on e-Social Science, Manchester, UK, June.

Blank, G. and Rasmussen, K.B. (2004)
"The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." Social Science Computer Review 22, no. 3 (August): 307-318.

Norwegian Social Science Data Services (1999)
"Providing Global Access to Distributed Data Through Metadata Standardisation: The Parallel Stories of NESSTAR and the DDI." Working Paper No. 10, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, September 22-24. (full text)

Ryssevik, J. (2000)
"Bazaar Style Metadata in the Age of the Web - An 'Open Source' Approach To Metadata Development." Working Paper No. 4, UN/ECE Work Session on Statistical Metadata, Washington, DC, November 28-30. (PDF 54K)

Miller, K. and Vardigan, M. (2005)
"How Initiative Benefits the Research Community - the Data Documentation Initiative" Paper presented at the First International Conference on e-Social Science, Manchester, UK, June.

Ryssevik, J. and Musgrave, S.(2001)
"The Social Science Dream Machine." Social Science Computer Review 19, no. 2 (summer): 163-174.

Green, A., Dionne, J. and Dennis, M. (1999)
Preserving the whole: A two-track approach to rescuing social science data and metadata (Technical Report 83). Washington, DC: Council on Library and Information Resources.

Bethlehem J, Kent J., Willeboordse A. and Ypma W. (1999)
"On the use of metadata in Statistical data processing", Working Paper No. 23, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999.

Nielsen, J. (1997)
"From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation". Unpublished thesis.

The Data Documentation Initiative Web site (http://www.icpsr.umich.edu/DDI/)

DDI Structural Reform Group (2005)
"DDI Version 3.0 Conceptual Model". Working draft by Gregory, A. and Thomas, W L.

About the Authors

Ken Miller (millk@essex.ac.uk) is the Information Systems Development Manager at the UK Data Archive (http://www.data-archive.ac.uk).

Ekkehard Mochmann is the Administrative Director of the Central Archive for Empirical Social Research, FRG. (http://www.gesis.org/en/za/index.htm).

Jostein Ryssevik is Senior Advisor at the Norwegian Social Science Data Services (http://www.nsd.uib.no/english/) and Technical Director of Nesstar Ltd. (http://www.nesstar.com).

Multi-mode and Multi-source Surveys

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Multi-Mode Research and Data Linkage. Theoretical and Practical Advice

George H. Terhanian

Abstract

Multi-mode research (specifically, research that involves multiple interviewing modes, often with different sampling frames and sampling methods) and data linkage (conceived broadly as the act of combining different surveys, projects, products, datasets and information systems, preferably through proactive design) are promising approaches that are likely to grow in importance with time. The aim here is to offer theoretical and practical advice on these approaches so as to inform science and society, and to improve research quality. Although the advice is directed primarily to market researchers, others might find it useful as well. It is intended to cross geographic boundaries, political and other jurisdictions, industry sectors, and academic disciplines

Keywords

Multi-mode research, data linkage,

1 Multi-Mode Research

No sampling frame is perfect. Nor is any data collection mode without limitations. For instance, the sampling frames associated with telephone and face-to-face interviewing modes, long the mainstays of commercial market researchers, are potted with holes. Telephone sampling frames often exclude, for example, individuals without land lines, students living in university housing, recent graduates residing in temporary housing, the incarcerated, the institutionalized, the hospitalized, those residing on military bases, and residents of countries such as Northern Ireland, states such as Alaska and Hawaii, and regions that include large sections of Brazil and Russia. In practice, they also exclude individuals who screen their calls or otherwise refuse to answer the telephone, people who travel or dine out frequently, those who work late, and individuals who rely heavily, if not exclusively, on their mobile phones. The sampling frames typically used in face-to-face survey research are incomplete as well. Individuals who reside in high-rise apartment buildings, certain gated communities, and those who are intensely private, among others, are effectively missing from the frame. The threats that these holes pose on research quality are likely to increase with time--the holes are getting bigger and potential respondents are becoming increasingly hard to reach.

Once researchers reach potential respondents, there is no guarantee that they will even participate in research. Unlike agricultural research in which "plots of ground can[not] excuse themselves from being treated" (Heckman, 1992, p. 215), it at least seems as though human beings are now excusing

themselves easily from participating in telephone and face-to-face research--refusal rates have been increasing steadily for years, particularly in the United States (US), thereby making these increasingly "hard to reach" respondents "hard to measure" as well (CMOR, 2004).

The difficulty of reaching "the hard to reach" and thereafter measuring "the hard to measure" through standard sampling and data collection approaches poses myriad challenges for researchers. The most passionate proponents of online research, always brimming with confidence, view these challenges as opportunities (Terhanian, 2003). Just as the automobile replaced the horse and buggy, the Internet will replace face-to-face and telephone interviewing, they say when they hawk their services. Unlike telephone or face-to-face respondents, these proponents add, online respondents are able to participate in research at their convenience, rather than that of the research organization. In practice, this means that those elusive respondents who screen their calls, choose not to answer the telephone, who dine out and travel frequently, work late, reside in university or temporary housing or high-rise apartment buildings, and use only their mobile phone are much easier to reach, and, consequently, much easier to measure as well. The ability to present images and video, to sample tiny portions of the population, such as gay or lesbian IT managers in small to mid-size firms, to reduce project costs and cycle time, and to eliminate interviewer effects is the icing on the cake for these passionate proponents of Internet research.

Their claims, characterized at times as shameless hucksterism in the late 1990's, are too important to dismiss today; one need only to glance at the growth rate of Internet research for compelling evidence of the seriousness of the enterprise. In 2005, market research organizations are likely to generate more than $1.3 billion in revenue through Internet research (Inside Research, 2005), remarkable growth from a starting point of $0 less than ten years ago. For the second consecutive as well, the European market is expected to grow at a faster rate than the US market, with annual growth of 43% expected versus 19% in the US. Doubtless, this will surprise and embarrass the more recent European skeptics of Internet research, who regarded the early growth as a unique American phenomenon (Jamieson, 2002). The sky is not falling for telephone and face-to-face research, however, partly because the sampling frames that Internet researchers depend on include their own fair share of holes: not everyone is online, those who are online do not necessarily belong to Internet panels (through which the majority of revenue is generated), and those who belong to panels, having typically joined through a non-random process, do not necessarily participate in surveys. Recent research suggests, as well, that some Internet users simply prefer to be interviewed by means other than the Internet (CMOR, 2004).

That the Internet may be an incomplete solution to the problem of reaching "hard to reach," "hard to measure" respondents should come as no surprise. The quality of Internet research, like the quality of telephone and face-to-face research, can vary dramatically—just as there are good and bad telephone and face-to-face surveys there are also good and bad Internet surveys. What does not vary among practitioners of high-quality Internet research, however, is a steadfast reliance on census information and periodic nationally representative surveys, both collected through traditional modes, as checks and means of possibly adjusting data collected online.

Absent these checks, it would be impossible to make population-wide inferences with any reasonable degree of confidence--it would be safer to place one's faith in tarot cards or a crystal ball. When these checks are in place and accompanied by thoughtful use of selection bias modeling approaches to compensate for coverage error, the effects of panel participation, and the non-use of probability-based sampling methods, however, Internet researchers might very well produce information that is superior to that which they might have produced through traditional approaches for many types of research studies, including pre-election polls (Terhanian et al., 2001). This kind of evidence notwithstanding, it

serves no useful purpose for researchers to squabble over the superiority of telephone, face-to-face or Internet interviewing. They are not in open warfare, let alone competition, and such a narrow view limits the possibilities that each mode offers.

In practice, the mode that researchers decide to use for most projects must depend on the characteristics of the target population and the area of inquiry, among other factors such as cost and cycle time. For example, to estimate the percentage of homeless people in Barcelona, the research organization might dispatch a team of face-to-face interviewers to the streets in a systematic manner. The estimates produced are likely to be far more trustworthy than those achieved through telephone or online research. Conversely, in order to understand why people in France or Germany are not online, it obviously makes more sense to employ a telephone-based approach, given the high telephone penetration rates in these countries. But to understand why Internet users, particularly those residing in countries such as China and Bulgaria with low Internet penetration rates, chose to use the Internet in the first place, an online approach would be the best choice inasmuch as the sampling frame would include only Internet users.

Neither in theory nor in practice, however, must researchers select only one mode. In many cases, quite possibly in most cases, it may be more prudent to employ a multi-mode approach. Through multi-mode interviewing, it becomes possible to plug the holes of each mode and produce samples that may be more fully representative of the target population--call this the scientific justification. The economic justification, which is somewhat narrower than the scientific one, is to reduce the cost per interview and the time required to complete these interviews, while the respondent-centric justification is to provide respondents with choice of mode in the interest of enhancing the survey experience and increasing response rates. Practically speaking, it is far easier to reach university students through email by sampling large online consumer panels than through telephone research that relies on population-wide probability sampling. Similarly, in some countries, it is also far easier to interview older women through telephone or face-to-face research than through Internet research. If one targets both groups as part of the same project, as one must often do, a multi-mode approach might be the ideal.

Designing good multi-mode studies is neither simple nor straightforward, and market researchers' understanding of multi-mode research, notably, the category involving multiple sampling frames and multiple sampling methods, is embarrassingly thin, almost too thin to consider a future in which multi-mode approaches are the rule rather than the exception. For example, imagine that a telecommunications company is interested in running a multi-country study of business decision makers. Typically, the market research organization might begin by purchasing a complete listing of business telephone numbers from one of its suppliers. It would thereafter make telephone calls to a representative sample of these telephone numbers during the work day in order to complete the desired number of interviews. As an alternative, it might consider a dual sampling frame, dual mode approach in which it relies on the listing of telephone numbers to complete interviews through the telephone or Internet (by directing potential respondents to an Internet-based survey) and a large Internet panel to complete interviews through the Internet. It might then attempt to complete as many as interviews as possible through the Internet to reduce the cost per interview and, possibly, the time required to complete these interviews. Uncertainty over how to put together the information collected through these modes would be the major stumbling block in this case. More generally, it is the primary reason why researchers shy away from employing such approaches. Practical concerns, such as the cost of project management (e.g., survey design, data collection, data processing), the need (usually) to use multiple interviewing systems, and the threat of a variety of possible mode effects (e.g., social

desirability bias) stand in the way as well. Even on those occasions when researchers do employ such approaches, they tend to depend on primitive methods of putting together the data, rather than more sophisticated methods that take into account the inherent differences between and among interviewing modes, sampling frames and sampling methods. Nevertheless, there is no reason, in principle, why data collected through multiple modes with different sampling frames and different sampling methods cannot be put together intelligently so as to enhance representativeness, improve cycle time and reduce cost. Looking to the future, researchers might have no choice other than to learn to mount these kinds of studies given the growing holes in traditional sampling frames, the limitations of some Internet-based frames, and ongoing demand for research. In light of this very realistic possibility, the next aim here is to offer advice on how researchers might design better multi-mode studies.

Advice on the Design of Multi-Mode Research

The prospect of more easily and affordably interviewing "hard to reach," "hard to measure" respondents is an appealing one. Intuitively, it also makes much sense that multi-mode, multi-sampling frame approaches could be more effective than those that depend only on one mode or one sampling frame inasmuch as the combination of modes and frames should be stronger than each one in isolation. The notion is neither pie in sky nor an unfamiliar one to market researchers. Today, in fact, many market research organizations, particularly those with a more global bent, conduct some types of multi-mode research.

To measure consumer attitudes and opinions in multiple countries, for example, they might conduct interviews by telephone where telephone penetration rates are high (e.g., Germany) and face-to-face where telephone penetration rates are low (e.g., Chile). Assuming that the sampling frames and sampling methods are similar across both countries, all that is required to make fair comparison between, say, Germans and Chileans, is an understanding of whether Germans would have provided the same responses to the questions they answered on the telephone if these questions had instead been asked through a face-to-face approach, and whether Chileans would have provided the same responses to the questions they answered in the face-to-face survey if the survey had been conducted by telephone. An understanding of cultural differences, and how these differences impact use of scales, is important as well. For example, Chileans might be more likely to offer high marks on customer service than their German counterparts, all else equal, simply because they use scales in a systematically different (and measurable) way.

Detecting mode effects and figuring out whether respondents from different countries use scales differently are less burdensome challenges than one might think, primarily because researchers have devoted serious thought to these kinds of issues through the years, often through controlled experiments in which survey modes have been randomly assigned to respondents in the interest of understanding whether question and response presentation (e.g., oral versus visual) and the source of presentation (e.g., interviewer versus self-administered) impacts the responses that are given. A deep understanding of the literature (e.g., see Dillman, 1978; 2000) will go a long way in reducing or minimizing mode effects in multi-mode surveys among different populations involving similar sampling frames and sampling methods.

Market research organizations also conduct multi-mode research among physicians (who are identifiable through medical associations) and the customers of its clients (in which case the client would provide contact information). In both cases, respondents would generally be sampled from one rather than multiple lists and given the option of completing the interview through a choice of modes.

To determine whether there might be mode effects, one need to assess whether the same person would have responded to the same question in the same way independent of mode. Through random assignment of survey mode to a sub-sample of respondents, or through use of a statistical substitute to random assignment such as propensity score adjustment (Rosenbaum & Rubin, 1983; 1984), it is possible to measure, and potentially adjust for, mode effects, as a precursor to combining data. In practice, it is straightforward to conduct this kind of study as well.

Conducting multi-mode research involving multiple sampling frames and multiple sampling methods (e.g., probability and non-probability) is less common, possibly because of uncertainty over how to deal with the sampling method problem. Neither the approach nor the problem is new, however. One can trace the theme to the mid-1950s. At the time, many prominent researchers had recently dismissed on methodological grounds the findings of the controversial *Kinsey Report on Sexual Behavior in the Human Male* (Kinsey, Pomeroy, and Martin, 1948). Of particular concern was the researchers' use of non-probability sampling to recruit and interview a sample of white, adult males through whom they attempted to make inferences about all members of this population. From a statistical standpoint, the notion that a sample selected through a process whereby respondents are not selected at random (i.e., a non-probability based approach) from the population universe offended the sensibilities of the report's critics. These critics simply refused to believe that the responses resulting from a convenience sample could fully represent the target population (Terhanian et al., 2001).

Prompted by these criticisms, the National Research Council invited the American Statistical Association to appoint a blue-ribbon committee to evaluate the Kinsey Report's methodology. The committee included three of the world's most able statisticians: William G. Cochran, John W. Tukey, and Frederick Mosteller. Unlike some of their colleagues from the statistical community, Cochran, Tukey, and Mosteller did not outright dismiss the Kinsey Report's findings. As they asserted: "... he did not use a probability sample is ... not a criticism which should end further discussion" (p. 328).

Recognizing that it is at times difficult or impossible to interview large numbers of respondents through probability-based sampling methods given the sensitive nature of certain topics, Cochran, Tukey, and Mosteller proposed as an alternative a dual sampling frame, dual sampling method approach to measure the sexual behavior of adult males. "Since it would not have been feasible for KPM to take a large sample on a probability basis," they observed, "a reasonable probability sample would be, and would have been, a small one and its purpose would be: (1) to act as a check on the large sample; and (2) possibly to serve as a basis for adjusting the results of the large sample" (p. 23).

Cochran, Tukey, and Mosteller did not offer more detailed advice on how Kinsey and his colleagues might have combined the two data sources, possibly because trustworthy tools for this purpose had not yet been invented. In fact, it would take another thirty years for Paul Rosenbaum and Don Rubin to do so.

Rosenbaum and Rubin (1983, 1984), colleagues of Mosteller's, developed a technique known as propensity score adjustment, a statistical matching tool designed to minimize the biases associated with the non-random selection of participants in observational studies. In other words, it is a selection-bias modeling technique that allows one to understand, and possibly adjust for, the differing reasons that motivate people to decide to engage in (i.e., to self-select) certain activities to which they have not been assigned at random, when one's interest lies in understanding the impact of these activities. For instance, to understand the impact of cigarette smoking on the incidence of lung cancer, one cannot randomly assign babies at birth to a lifetime of smoking, or non-smoking, for ethical and practical reasons. Propensity score adjustment, in this case, would allow one to make fair comparisons between

smokers and non-smokers in the absence of random assignment so as to produce a credible estimate of the effects of smoking.

Applying tools such as propensity scoring in a thoughtful manner to improve multi-mode survey design is another matter altogether, primarily because these tools have typically been used opportunistically to re-analyze (or meta-analyze) existing data sets. For instance, the US General Accounting Office (USGAO, 1995) relied partly on propensity scoring to combine data generated on cancer patients through database analysis with that generated through small-scale controlled experiments. They did so without the benefit of having designed these data sources to permit linkage in the first place, thereby reducing the insights that could be gleaned.

In order to invert the analytic focus of tools like propensity scoring to improve multi-mode survey design involving different sampling methods, one requires, at the very least, overlap between sub-samples of respondents that are interviewed via different modes and sampled through different methods (Boruch & Terhanian, 1996, 1998; Terhanian & Boruch, 2000; Terhanian et al., 2001). Ideally, one sub-sample will be the product of a probability-based sampling method, although this is not mandatory. More important, the combined sample must represent the target population for the approach to have any value at all.

To illustrate further, consider the multi-mode approach that Harris Interactive has employed to contend with the many obstacles, notably, coverage error, self-selection, and the effects of panel participation, associated with making population inferences through Internet based research. The approach (Terhanian et al., 2001) is akin to the solution that Cochran, Tukey and Mosteller described more than fifty years ago.

To begin with, Harris relies heavily on its Internet panel of several million potential respondents as one sample source. When it does so, it recognizes that it must take into account at least three important decisions that panelists have made.

First, they have decided to become part of the online population. For many, this decision is a function of the costs of a computer and Internet access. Second, members of the online population have decided to register for and join the Harris Poll Online (hereafter HPOL), the research panel the company maintains. Third, members of the HPOL have decided to respond to an invitation to complete a survey.

The decision to respond to a survey invitation, however, is preceded by two other important decisions. Because a person must first be online to join the HPOL and because Harris sends survey invitations only to those who have joined, the Internet-based samples do not contain information about those who are not yet online or about those who are online but have not yet joined the HPOL.

For these reasons, from time to time (typically, monthly), Harris conducts parallel (same questions at the same time) telephone or face-to-face and online surveys. The aim of these surveys is to compare a representative sample of the population of the US, the United Kingdom (UK), France, Germany and other countries of interest with a representative sample of its panel members, most of whom have been recruited by means other than probability sampling.

After Harris completes the interviews through these parallel modes, it employs logistic regression to develop a statistical model that estimates the probability that each respondent, conditional on his or her characteristics, participated in the telephone or face-to-face study rather than the Internet one. The probability, or "estimated propensity score," is based on answers to several socio-demographic, behavioral, and attitudinal questions. Next, in the "propensity score adjustment" step, Harris groups

respondents by propensity score within the survey group (telephone/face-to-face or Internet) they represent. Statistical theory suggests that when the propensity score groupings are developed methodically, the distribution of characteristics within each Internet grouping will be asymptotically the same as the distribution within each corresponding telephone or face-to-face grouping. Therefore, by weighting the Internet sample's propensity group proportions to be the same as the telephone or face-to-face sample's proportions, the distribution of characteristics will the same across all groupings.

This procedure produces a result similar to randomization: the estimated probability of belonging to one group rather than the other will be the same given the variables in the model. And because the model includes behavioral, attitudinal, and socio-demographic information, Harris is far more confident that it has reduced or eliminated bias than it would have been if it had relied only on basic socio-demographic variables. More important, there should be no differences in the survey responses of interest to its clients irrespective of mode, as long as the propensity score model includes the right variables and the survey has been designed in a way to minimize mode effects.

For subsequent Internet-based surveys, Harris estimates each respondent's propensity score using a model it developed earlier in the month; a carefully constructed propensity model can be ported across surveys inasmuch as the unit of measurement is the respondent rather than the response.

The multi-mode approach described here has allowed Harris to make population-wide inferences in many subject areas, and the accuracy of these inferences has been well-documented (e.g., see Chiang and Krosnick, 2001; Berrens et al., 2003). It has also allowed Harris to mount additional kinds of multi-mode surveys with scientific rigor and confidence. For instance, to explore issues of personal privacy, one might mount a multi-mode study in which a small sample of respondents is interviewed by telephone or face-to-face and a larger sample of respondents is interviewed via the Internet. The rationale for such an approach is at least three-fold:

1. The client and the market research organization suspect that the views of Internet and non-Internet users will differ on matters of personal privacy. For this reason, it is necessary to interview non-Internet users.
2. An Internet-based approach allows one to explore Internet-specific issues in greater detail (e.g., by directing respondents to web-sites during the course of the survey) than telephone or face-to-face approaches.
3. The Internet offers additional advantages such as lower cost per interview and speed.

The major challenge here would be to figure out whether and how to pool the responses of the Internet users who are interviewed via different modes. Inasmuch as Internet users that join Internet panels may be less concerned about their personal privacy than their counterparts who do not join Internet panels, all else equal, it makes sense to test the hypothesis prior to pooling; the data collected from Internet users through the non-Internet survey would serve as a check, and as a means of possibly adjusting the data collected online. The test would increase one's confidence in the credibility of the responses once the data have been put together.

In summary, if a sub-sample of respondents with the same characteristics answers the same questions the same way irrespective of mode, sampling frame and sampling method, then it is often reasonable to conclude that data collected through multiple modes can be put together in a way that might enhance sample representativeness while reducing cost and field time (Terhanian et al., 2001).

Tools such as propensity scoring, when employed thoughtfully, promote and facilitate multi-mode research in at least three ways: (1) by serving as an enhancement to, or substitute for, random assignment when estimating mode effects, (2) by enabling researchers to eliminate or reduce socio-

demographic, attitudinal, and behavioral differences between and among modes, and between probability and probability and non-probability samples, and (3) by allowing researchers to make population-wide inferences, at times, based on the responses of individuals sampled by means other than population-wide probability approaches, irrespective of interviewing mode.

2 Data Linkage

The astronomical growth of Internet research is certainly not the only interesting story in market research in the past decade. During these years, the research community has also been exposed to a flurry of analytic advances that promote and facilitate data linkage (Terhanian et al., 2004), conceived broadly as the act of combining different surveys, projects, products, datasets and information systems, preferably through proactive design. It is not clear, however, that researchers, or their clients, understand fully the implications of these advances. In other words, it is not yet commonplace for them to regard every study as a piece of a more complex puzzle (Terhanian and Boruch, 2000). Even among those that take seriously the "puzzle pieces" view, questions continue to abound on how to put together pieces that have not been designed to be put together. In practice, the task can be a daunting one, irrespective of the availability of analytic advances that permit data synthesis (e.g., meta-analysis) and fusion. Impeding the process are obstacles such as incompatible sampling frames, questions asked somewhat differently across studies, differing dates of data collection, and organizational inefficiency. In short, those who attempt to put together these puzzle pieces may at times come to the realization that "certain pieces seem broken, several duplicate pieces exist, some pieces are inexplicably missing, and a few new pieces are produced so slowly that they appear to be altogether lost" (Terhanian and Boruch, 2000. p. 218).

In the US, for instance, the National Center for Education Statistics (NCES) periodically conducts separate surveys of students and teachers from within the same school. In principle, it should be straightforward to modify sampling plans, tinker with question wording, and the like so as to link student, teacher and school data in the interest of identifying new insights on how students learn. In practice, this has not happened, partly because the architects of these surveys have not worked together closely. As one consequence, there is minimal overlap among the various NCES data sets, making them unnecessarily unlinkable (Terhanian and Boruch, 2000).

At other times, it may at least seem possible to link together data via common variables such as age, gender, or education level, but linkage may not produce a more meaningful data set. At worst, it might even lead to false insights. A young, highly educated female sampled through a traditional probability-based method such as telephone might possess similar views on one issue as a young, highly educated female sampled by other means (e.g., Internet). This does not mean, however, that these women will possess the same views on all other issues and consequently that their data can be merged together. To make this assumption is risky business without any additional information; that is, it requires a high tolerance of a high degree of uncertainty (Terhanian et al., 2004).

Researchers who attempt to employ linkage tools (no matter how statistically sophisticated these tools may be) on data that are not designed to be linked together run into these kinds of problems all the time. The simple implication is that researchers ought to ask themselves how they can design better surveys, projects, and information systems to promote and facilitate linkage. The next section offers advice on how they might do so.

Advice on How to Promote and Facilitate Data Linkage

The idea of putting together data from different sources in the interest of producing or educing deeper insights (or at least increasing the use and usefulness of existing data) can be traced at least as far back as John Graunt's efforts in the 17th century to understand how to combine data in the interest of the crown. Graunt encouraged the crown to learn about the kingdom through multiple sources of information, including counts of soldiers-at-arms, recorded numbers of births and deaths and many other sources (Terhanian & Boruch, 2000). Graunt did not make plain how he combined these data, however.

Alexander Graham Bell also capitalized on the notion of linkage in his study of genetic transmission of deafness. He depended in the late 1880s on completed Census Bureau interview forms found strewn in a government building basement and on genealogical records from other sources (Bruce, 1973). It is not clear how he combined these information sources either.

Similarly, atmospheric weather researchers worldwide rely partly on satellite imaging to measure large grids from space and partly on ground-based systems to obtain measures on smaller grids. As with Graunt and Bell, however, it is not clear how they combine the information (Draper et al., 1992).

In theory, the task of linking data together hinges on enhancing the extent to which major factors are common to different surveys, projects, datasets and information systems. It also depends on creating ways to induce artificial commonness. More practically, in order to promote and facilitate linkage, it is helpful to begin by thinking of all possible data sources as pieces of a large puzzle. One might continue by selecting as primary one study, data set, or information source. When multiple studies are equally important, one should arbitrarily designate only one as primary. Any linkage between this primary study and other studies or datasets should then be viewed as a linkage that involves augmentation of the primary study. The next challenge might be to figure out how to augment the primary data. The following types of questions might serve as a starting point (Terhanian and Boruch, 2000):

- What variables are common to various data sources?
- What ways of measuring each variable, ways of sampling, and administration are common, making linkage among datasets easy?
- What differences in ways of measuring, administrating, and sampling make linkages difficult?
- What can be done to modify different datasets so they are linkable in some way?

It might then make sense to define the kinds of linkage that one might consider (Terhanian & Boruch, 2000). As an example, consider the eight that follow:

1. Variable Augmentation: New variables, generated by different sources and observed on the primary sample, are added to the primary sample dataset based on exact or statistical matching approaches. For instance, one might add contextual variables bearing on the area in which respondents reside based on an exact match on post code. Or one might link personnel records bearing on exposure to possibly dangerous chemicals to the same individual's death records to understand whether certain chemicals might be associated with premature deaths. Statistical matching, which aims to link individuals, or other units or entities, based on socio-demographic or other characteristics, expands the opportunities that are available by linking like with like, rather than exact with exact. For example, Harris Interactive and CBS Television (Terhanian et al., 2004) developed an approach that exploited three information sources--the Nielsen Media Research (NMR) national people meter (NPM), Harris Poll telephone research, and Harris Poll online research--to estimate out-of-home viewership of those included within the NPM sampling frame. In short, Harris and CBS statistically matched respondents in the NPM panel to those who participated in Harris Poll research based on socio-demographic and other characteristics in order

to add variables on out-of-home viewing to the NPM base case—NPM measures only within-home viewing. The apparent success of the approach augurs well for multi-mode research and data linkage, individually and in combination.

2. Sample Augmentation: A different sample of the same target population is put together with the primary sample, to keep costs down or to enhance or deepen understanding. On many occasions, for example, the high cost of random-digit-dial telephone research dissuades clients interested in understanding rare populations from commissioning such research. Consider a client interested in surveying a nationally representative sample of 2,000 US adults, 1,000 of whom are African-American (Terhanian et al., 2001). Because African-Americans constitute approximately 12% of the adult population in the US, using a pure random-digit-dial approach to survey 2,000 US adults would yield approximately 240 African Americans. Consequently, this approach would leave the client short by 760 African Americans (1,000 required – 240 obtained). To obtain the additional 760 African Americans would require the organization to survey an additional 6,334 US adults (6,334*.12 = 760), bringing the total number of adults surveyed up to 8,334 (2,000 + 6,334). To reduce the cost of the project, the research organization might propose purchasing a list of telephone numbers of known African-Americans built by means other than census or probability sampling. It might then use the RDD approach to interview a random sample of 2,000 adults, of whom 12% or 240 would be African-American, while using the purchased list to contact the additional 760 African-Americans. To put together the two samples of responses of African-Americans, the research organization might exploit a propensity score approach. Conceptually, the sample of 240 African-Americans interviewed via RDD would be regarded as one group, the sample of African-Americans interviewed based on the imperfect list would be regarded as the second group, and differences between the two groups would be modelled through logistic regression. Adjustment based on the propensity score would then follow as a precursor to putting together the data. The approach described would reduce the overall cost of the project without necessarily jeopardizing the generalizability of the study's findings. It would also enhance the data set on African-Americans via sample augmentation.
3. Time Augmentation: New measures are put together with earlier measures of the same variables on the same sample, as in a longitudinal study that tracks growth over time.
4. Family Augmentation: Measures taken from family or friends of units in the primary sample are added to the primary sample. For instance, data from "best friends" or "peers" might be added to student data to understand the impact of these relationships on activities like smoking and drinking, with implications on how to curtail these unhealthy practices.
5. Levels Augmentation: Measures taken on units at a higher level than the units in the primary sample are added to the primary sample dataset. For example, nation- or region-level policy variables may be observed and added to individual records in multi-country studies. Or school district, school, or classroom might be added to individual student data in a study to understand the factors that drive student achievement.
6. Mode Augmentation: New ways of measuring roughly the same variables on roughly the same units are added to a primary sample dataset. For instance, digitized videotape data may be added to teacher and student records in the same schools in two countries as a different way of measuring what is taught and how.
7. Population Augmentation: New populations having been surveyed using the same measures are put together in a file with the primary sample dataset, the primary sample having been drawn from a different population. For instance, a new, Norwegian sample might be added to a multi-country data set that heretofore had no Norwegian data.
8. Replicative Augmentation: A different sample of a different or the same population using identical measures is put together with a primary sample dataset, often as a check on the credibility of the primary sample.

The listing of linkages described above is intended to be illustrative rather than exhaustive. As in the past, it continues to make sense to employ a simple tool to make them memorable, no matter how miserable the music might still be (Terhanian & Boruch, 2000, p. 207):

Recall the musical scale: doh re me fah soh lah te doh? Change a couple of letters and we get:

- Voh: Variable augmentation
- Re: Replication augmentation
- Me: Mode augmentation
- Fah: Family augmentation
- Soh: Sample augmentation
- Lah: Level augmentation
- Te: Time augmentation
- Po: Population augmentation

More seriously, no shortage of opportunity exists for market research organizations to take the lead in designing surveys, products, projects and datasets in a way that promotes linkage between and among them, and to other data sources. Arguably, organizations that do so will gain competitive advantage that will reflect and reward the quality of thinking and effort involved in extending the puzzle pieces metaphor to its fullest by regarding every study, every variable within each study, and every dataset as linkable through exact or statistical matching. Indeed, in some ways, this represents a timely opportunity for market researchers to shatter any stereotypes, deserved or undeserved, of the level of value they bring to clients, always a perennial concern. Although many market research organizations trumpet their consultative capabilities, clients and prospective clients at times prefer to look elsewhere. "We'd like you to collect the data but our consultants will take the lead in educing the implications of these data," is something market research organizations may hear far too often. If market researchers adopt a more aggressive posture toward data linkage, get much better at it, and think more deeply about the implications of these data on their clients' business needs, they will hear this less often.

3 Final Thoughts and Conclusions

Multi-mode research involving multiple sampling frames and sampling methods need not be the knotty conundrum that it is for market researchers and others. By tapping into the growing literature on mode effects (e.g., see Thomas, 2004; Taylor, 2005) and thinking creatively about selection biases and how to overcome them, it is possible today to take advantage of the many benefits of multi-mode approaches.

Although discussion here focused heavily on telephone, face-to-face and Internet data collection modes, it will be important in the future to investigate additional modes for their potential contribution to multi-mode research. In principle, there is no reason why data collected via telephone, face-to-face, Internet, fax, mail, people meter, diary, and hand-held devices, among others that may not yet have been invented, cannot be put together in a way that benefits from the strengths of each mode without suffering from the limitations. Developing an understanding of these strengths and limitations (and how to overcome them) remains a work in progress, although the availability and thoughtful application of selection bias-modeling tools such as propensity scoring coupled with creativity and determination will make this work less difficult. In parallel, the continued development of software to facilitate multi-mode data collection will only increase demand for such approaches, inasmuch as more flexible software will reduce or eliminate one more obstacle.

The discussion here on multi-mode research is not unrelated to that on data linkage. Tools such as propensity scoring are not only instrumental in conducting good multi-mode research, but they also pave the way in promoting and facilitating data linkage through statistical matching. It makes no

sense, however, to rely only on statistical matching to link together data, or to re-analyze existing data, as some have done in the past. It is much easier to design studies, products, datasets and any other sort of information system that might inform policy and practice with linkage in mind. The eight ways to link data, described earlier, provide a starting point. But they are only a starting point, and market researchers, and their clients, have a long road to travel. Today, for example, the same client might commission multiple independent surveys, or purchase multiple independent datasets or products, from the same research organization but neither the client nor the research organization will have any idea of how they might link this information together, let alone to other data sources. As one consequence, the results of these puzzle pieces cannot be easily integrated, compared, or combined, no matter how hard one might try.

Multi-mode research and data linkage offer a river of possible benefits. Armed with a deeper understanding of how to mount multi-mode studies and how to design surveys, products, projects, and datasets in a way that facilitates linkage, market researchers may be able to increase their contribution to science and society, strengthen their relationships with their clients, and enhance their stature within the wider community of researchers. It is time to let the river flow.

References

Berrens, R.P, Bohara, A.K, Jenkins-Smith, H., Silva, C., and Weimer, D.L (2003).
The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples. Political Analysis, 11:1. Society for Political Methodology.

Boruch, R.F., and Terhanian, G. (1996).
So What? The Implications of New Analytic Methods for Designing NCES Surveys. In G. Hoachlander, J.E. Griffith, and J.H. Ralph (Eds.), From Data to Information: New Directions for the National Center for Education Statistics, NCES 96-901, Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Boruch, R.F. and Terhanian, G. (1998).
Controlled Experiments and Survey-Based Approaches to Productivity Research: Cross Design Synthesis. In H. Walberg and A. Reynolds (Eds.), Advances in Educational Productivity. Greenwich, CT: Jai Press.

Bruce, R.V., Bushery, J., Royce, D. and Kasprzyk, D. (1973).
Bell: Alexander Graham Bell and the Conquest of Solitude. New York: Little Brown.

Chang, L. and Krosnick, J.A. (2001).
RDD Telephone vs. Internet Survey Methodology: Comparing Sample Representativeness and Response Quality. Paper presented at the 56th Annual Conference of AAPOR, Montreal.

CMOR (2004).
Respondent Cooperation and Industry Image Survey. Report of the Council for Marketing and Opinion Research. Respondent Cooperation Committee's Study on cooperation levels.

Cochran, W.G., Tukey, J.W., and Mosteller, F.M.
(1954). Statistical Problems of the Kinsey Report. Washington, DC: The American Statistical Association.

Dillman D.A. (1978).
Mail and Telephone Surveys. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2000).
Mail and Internet surveys: the Tailored Design Methods. (Second edition). New York: John Wiley & Sons, Inc.

Draper, D., Graves, D. P., Goel, P. K., Greenhouse, J., Hedges, L. V., Morris, C. N., Tucker, J. R., and Waternaux, C. M. (1992)
Combining Information for Research. Washington, D.C.: National Academy of Sciences. (Also in: Contemporary Statistics. No.1, Alexandria, VA: American Statistical Association, Undated).

Heckman, J. (1992).
Randomization and Social Policy Evaluation. In Charles F. Manski and Irwin Garfinkel (Eds.), Evaluating Welfare and Training Programs. Cambridge, MA: Harvard University Press.

Inside Research (2005)
Issue 203: 1-2.

Jamieson, D. (2002)
Online Research Gets Few Euro Cheers. Marketing News, January 21: 15.

Kinsey, A.C., Pomeroy, W.B., and Martin, C.E. (1948)
Sexual Behavior in the Human Male. Philadelphia: W.B. Saunders.

Lippman, W. (1963)
"The Savannah Speech". In C. Rossiter & J. Lare (Eds.), The Essential Lippman. New York: Random House. [Originally published in 1933]

Rosenbaum, P.R., and Rubin, D.B. (1983)
The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41-55.

Rosenbaum, P.R., and Rubin, D.B. (1984)
Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association, 79 (387): 516-524.

Taylor, H., Krane, D., and Thomas, R. K. (2005)
How Does Social Desirability Affect Responses? Differences in Telephone and Online Surveys. Paper presented at the 60th Annual Conference of the American Association for Public Opinion Research in Miami Beach, Florida.

Terhanian, G. (2003)
The Unfulfilled Promise of Internet Research. Market Research Society Conference, Paper 37.

Terhanian, G., and Boruch, R.F. (2000)
Putting Surveys, Studies, and Datasets Together: Linking NCES Surveys to One Another and to Datasets from Other Sources. In the National Research Council's Grading the Nation's Report Card, Washington, DC: National Academy Press.

Terhanian, G., Bremer, J., Delaney, T.F. and Thomas, R.K. (2004)
Measuring Television Viewership through a Multi-Method Approach. Proceedings from ESOMAR/ARF Worldwide Audience Measurement Conference. 183-198.

Terhanian, G., Smith, R., Bremer, J., and Thomas, R.K. (2001)
Exploiting Analytical Advances: Minimizing the Biases Associated with Non-Random Samples of Internet Users. Proceedings from ESOMAR/ARF Worldwide Audience Measurement Conference. 247-272.

Thomas, R. K., Krane, D., and Taylor, H. (2004).
On the Convergent Validity of Attitude Measurement in Phone and Online Surveys. Paper presented at the 59th Annual Conference of the American Association for Public Opinion Research in Phoenix, Arizona.

U.S. General Accounting Office (1995).
Breast Conservation versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies. Washington, D.C.: U.S. General Accounting Office.

About the Author

George H. Terhanian, Ph.D., is President, European Operations & Global Internet Research, Harris Interactive.

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Adding Value to Data Through Improved Access. The Case for Web Portals

Reginald Baker

Abstract

With the increasingly widespread use of the Internet has come a host of new opportunities to leverage the value of survey data. By creating broader access and disseminating easy-to-use analytical tools we not only can make more data available to a broader set of users, we also can minimize the barriers that might exist among the discrete activities of data collection, data analysis, report development, and delivery. One prime example in the market research world is the deployment of Web portals designed to deliver data more quickly, to increase the value of those data by making them easier to analyze, to deliver data and results deeper into the client's organization, and to retain historical data for future analyses in an online archive. This paper describes a case study in the development and evolution of one such portal from an online document store to survey progress tracker to analysis and reporting platform to data archive. It also looks to the future and speculates on the kinds of features we might expect as these systems increase in complexity and sophistication.

Keywords

Web; Internet; Online Analysis; Reporting; Metadata; Paradata

1 Adding Value in Survey Research (The Context)

The application of information technology (IT) to survey research arguably has been the most compelling issue facing our industry over the last 20 years. We are, after all, an industry that collects, manages, and analyzes information. It seems obvious that better, smarter, and more effective use of IT should be a constant goal.

The Information Value Chain

Porter's concept of the value chain (1985) offers a useful way for us to conceptualize the benefits of applying IT to the sequence of activities that begins with data collection and ends with analysis and reporting. Many of the advantages of IT described by Porter—shorter cycle times, lower costs, strategic links to suppliers and customers, and improved integration of internal processes—have been clear goals of the amalgam of technologies that includes CATI, CAPI, CASI, and, more recently, CAWI. Blattberg, Glazer, and Little (1994) have expanded on Porter's concept to propose the information value chain as a way of understanding IT's impact on market research (MR). Their value

chain (similar to Figure 1 – next page) describes the process by which raw data are collected and transformed into information for use in analysis and decision-making.

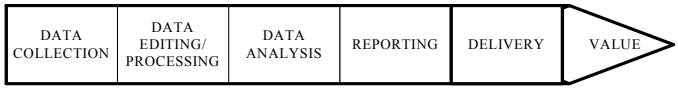


Figure 1: The Information Value Chain for a Survey Organization

At each stage of the process the survey organization processes data from the previous stage to add value through activities such as editing raw data, coding open ends, merging with other data, creating composite variables, performing analyses, or documenting and packaging for analysis. Put more simply, the overall goal is to these raw data and make them better by eliminating defects, preparing them for use, making them available sooner, explaining what they say, or making them easier to access and use. Within the context of the economics of the firm, the goal is to produce a product (data) whose perceived value is greater than the cost of producing it. In MR the value perception comes down to what customers will pay. For government and academic data where judgements about value may be more abstract the calculus is somewhat different, but the underlying principle the same.

Recent Developments

Throughout most of the late 1980s and 1990s our industry focused on the development of the so-called CASIC technologies (Couper et al., 1998) with the goal of adding value through greater accuracy, faster turnaround, and the ability to support increased design complexity. The major thrust of these improvements was to combine a number of formerly discrete steps (interviewing, data conversion, and editing) into a single unified process. Beginning in the late 1990s we began to appreciate the value of richer information about the data itself (metadata) and the data collection process (paradata). The emphasis on the former has significantly enriched publicly available data, while the latter continues to be of great value not only as a form of documentation for individual datasets but also as input to survey improvement activities (i.e., adding value) in general (See, for example, Jeavons (1999) or Couper (2000)).

Finally, we come to the present day, the "Internet Age," where the value of the Web as a survey data collection platform may soon be eclipsed by its value as a platform for survey data dissemination and analysis. The use of the Web as a dissemination vehicle for government statistics and datasets is already old news as services such as those offered by the Bureau of the Census in the US and the Office of National Statistics in the UK clearly demonstrate. Private archives such as the Interuniversity Consortium for Social and Political Research and the Roper Center have moved significant portions of their holdings online. GSSDIRS (http://webapp.icpsr.umich.edu/GSS/) offers an example of a single, ongoing study that disseminates data to the academic community via the Web.

In MR, Web-based interviewing systems have emerged with integrated sample management and analysis systems that provide access to data in real-time, that is, during data collection. Systems such as Confirmit from FIRM (http://www.confirmit.com/) and Net-MR from Global Market Insite (http://www.gmi-mr.com/en/net-mr/net-mr_software.phtml) include integrated reporting portals with basic analysis and graphing capabilities designed to reduce the cycle time and economize effort associated with analyzing and reporting on survey data collections. In some of these systems, intuitive

interfaces help users build self-documenting composite variables, case filters, weights, and report formats as well as share the results of their analysis with other users.

The Opportunity

The thesis of this paper is that the current greatest opportunity for delivering additional value to survey data users is at the end of the value chain, that is, the point of delivery. Where it was once sufficient to deliver a dataset with documentation we can now add significant value by bundling those data with the tools to manage, analyze and report on them. By hosting this activity in real-time via a portal on the Internet we also can bring survey data users deeper into the survey data development process. We can create a partnership among those involved in survey design, survey execution, and survey analysis that in the end produces survey data that are more transparent to users and better suited to their use.

The remainder of this paper first describes the development of one such portal, MSI's MSIClient.net service. It charts the portal's evolution from online document store to survey progress tracker to analysis and reporting platform to data archive. It then looks to the future and speculates on the kinds of features we might expect as these systems increase in complexity and sophistication. The overarching theme is the continued identification of opportunities to leverage technology in ways that increase the value of survey data through an enlarged set of benefits to survey data users. Sometimes those benefits come as reduced effort, sometimes in timelier reporting, and sometimes in helping us do things that we simply could not do before.

2 Case Study: The Evolution of MSIClient.net

Our story begins in 1998 with the desire of a researcher to share survey results with data users (clients) quickly, securely, and across a wide variety of locations around the US. Email had been tried and found wanting in required effort, accessibility, and security. The Web was an obvious solution.

Clients were slow to embrace the approach with concerns about Internet security and a general lack of comfort with Web-based interaction. But with time clients saw the value and became the driving force in its development demanding more features, broader access, improved ease of use, greater security, and increased customization. The direction was always clear: create a 360-degree view of survey data development and disseminate survey results faster and deeper into the client's organization.

Online Document Store

Our starting point was the development of an application we call *Documents* where electronic documents of all kinds can be posted and retrieved by anyone with access to the portal. These include proposals, research designs, questionnaire drafts, sample specifications and files, tabulations, and reports. This information store provides a working space for the key players inside and outside of the research organization during survey development. It ensures that everyone has the same information available at all times and reduces the potential for confusion about questionnaire versions, data collection schedules, sample specifications, and so forth. Once data collection is complete it becomes an invaluable historical resource of critical metadata needed to inform data analysis and use. It contains not just the questionnaire, but also the documents that can help the data user understand the analytic purpose behind the questions and their evolution from initial design through final questionnaire. In that sense, it extends the metadata concept by documenting the context of survey

design and development. It is a "self-documenting" feature, providing a depth and richness of documentation that often is lacking in traditional data archives.

Progress Reporting

Once clients became accustomed to interacting with us over the Web new opportunities surfaced. Survey sponsors have a natural interest in following the progress of data collection and disposition of the sample. Survey managers share this interest and have an equally compelling interest in timely measures of the cost of the effort. Reporting these data has been a consistent goal of CASIC (See, for example, Baker (1987) and Connett (1998))

The systems to generate these data and reports already existed within the company. The challenge was to provide a place to bring them together in ways that were informative and accessible both by survey managers inside the company and our clients. While the majority of our work continues to be telephone, there is a growing proportion of Web (approaching 25 percent) and a small (less than five percent) component of mail. On MSIClient.net we have combined these streams into a single reporting system that provides an overview of the entire data collection effort. It offers breakdowns by mode, by sample strata, and by any number of sample characteristics and combinations thereof. These reports are updated approximately every hour for Web, at the end of a shift in the telephone center, and as mail questionnaires arrive and are processed through keying or scanning. Our survey managers and their clients always have an up to date view of progress in data collection.

The point of providing this information is to create a capacity to act. The goal is not to be able to watch the train wreck in real-time, but to prevent it. Heeringa and Groves (2004) describe a process they call "responsive design" that is only possible in the real-time, information-rich environment created by systems like MSIClient.net. Heeringa and Groves expand the traditional concepts of the field test and sample reps into a breakdown of data collection itself into discrete phases, each designed to achieve specific survey outcomes. As a phase is completed the survey managers analyze sample dispositions, data collection costs, and response data. They use this analysis to adjust the design parameters of future phases to optimize across the standard measures of data quality, timeliness, and cost. Responsive design seeks to maximize survey value by spending money wisely and by ensuring that we get the best possible survey dataset given the money and time available to collect it. We can only do responsive design where we have real-time data on survey progress, costs, and quality outcomes.

Analysis and Data Delivery

While there is obvious value in the capabilities brought by an online document store and real-time progress reporting the analysis features of MSIClient.net and similar systems generally attract the most interest. Researchers are quick to appreciate the ability to analyze data during data collection, to quickly and easily generate reports, and to have someone else assume the burden of maintaining datasets online for extended periods. Such systems also can provide unprecedented access to data by data consumers who may not be researchers (e.g., executives, line managers, journalists, librarians, or just ordinary folk). If properly designed, they should have no special requirements beyond an Internet connection and a Web browser.

Accessing survey data over the Web is hardly a new idea. At one end of the spectrum are the Web services of national statistical agencies such as ONS, Statistics Canada, Statistics Netherlands, and the

US Bureau of the Census. These systems provide public access to huge datasets via simple interfaces. Where once users of such data had to go to libraries and wade through volumes of printed reports in the hope that the desired data tabulation had been created we can now generate much of what we need from our offices or our homes in minutes. At the other end of the spectrum are sophisticated, online analysis tools that are directed primarily at the skilled researcher. For example, web-based tools such as Harmoni (http://www.infotools.com/harmoni/index.htm) and Vector (http://www.cobalt-sky.com/products/vector/default.htm) offer capabilities that begin to compete with some offline systems (e.g., SPSS and SAS), at least in terms of the analytical tools that researchers use most often. The online systems generally do not offer the variety and sophistication of tools available in offline systems and their data management capabilities are significantly less powerful, but they nonetheless offer a compelling value proposition. They relieve researchers of the arduous task of acquiring and installing analytic software or buying data and building analytic datasets. For data generators, they reduce the cost of preparing and publishing analytic datasets, especially where analysis systems can interface directly to data collection systems via components such as the SPSS MR Data Model (http://www.spss.com/DDL/dm.htm) or where the analysis component is an extension of the data collection system as with Pulsar (http://www.pulsetrain.com/solutions/application/pulsar.htm) or NET-MR (http://www.gmi-mr.com/en/net-mr/reporting-item3.phtml).

The thrust of MSIClient.net development has been to add value by delivering data analysis to data consumers with varying analytical skills and a primary interest in descriptive analysis and reporting. A key design goal has been easy-to-use functionality that produces easy-to-understand outputs. We strongly believe in the value of putting data into the hands of people who need them to guide action whether they be line managers in a commercial firm, directors in a public agency, university researchers, or simply people whose professional or personal lives are enriched through more and better information.

The basic requirements that derive from these design goals have seemed to us to be (1) an intuitive, self-documenting interface for selecting data items and (2) outputs in formats for reporting tools that users already know how to use. Our solution to the first requirement has been to offer simple pick lists with variable labels and one click access to full question text. In the case of created variables such as recoded items users can review the recode logic. With a modicum of additional effort users can view the entire questionnaire to understand the context within which the question was asked. Our solution to the second is to offer options to export analysis tables into the standard components of Microsoft Office (Word, Excel, and PowerPoint).

The analytic tools available through MSIClient.net and similar portals developed by other firms tend to be pretty basic: frequencies, crosstabs, means, and the ability to sort and display open end responses by combinations of closed end responses. Users can recode variables, test for statistical significance, build and save filters to select subsets, and save report definitions to rerun or modify in a later session. Tables can be exported to PowerPoint into a user-supplied template that incorporates his or her personal or corporate standards. Users wishing to do more complex analysis can easily build custom extracts of cases and variables for export to Excel or SPSS. This basic functionality supports much of the analysis and report writing done by researchers inside the company, researchers in client organizations, and even some individuals in client organizations who are not professional researchers.

3 Beyond the Basics

Portal functionality of the sort described in the previous section is becoming a "must have" in the rapidly evolving world of MR. But clients continually demand more and research companies must always be looking for ways to deliver additional value.

Meeting the Needs of the Non-Technical Data Consumer

More and more Web portals are focusing on delivering data in ways that are useful irrespective of the analytical sophistication of the data consumer. Most contain the typical analytical tools needed to meet the majority of the day-to-day needs of professional market researchers both inside the research organization and in clients' organizations. Users with more sophisticated analytic needs can design their own extracts and download them for use in the major statistical packages. There are, however, large numbers of people, probably the clear majority of people, who lack the skill and the interest to do basic statistical analysis but nonetheless find value in survey data. In the MR world these are directors, managers, and supervisors in client organizations.

An increasingly important goal for portal developers is the delivery of survey data to the desktops of this group of users in forms they can use and in real-time. In general, two types of reporting seem to be favoured. The first is dashboard reporting that includes summaries of the key measures of interest, often trended over time and compared against agreed-upon goals and presented in graphical formats. Each time a user goes to the dashboard it is refreshed with the most recent data. The second is targeted reports specific to the management responsibilities of individual users. These are designed to cut the data in two ways: vertically by variable or measure and horizontally by span of responsibilities. For example, a manager may see a report that focuses on a restricted set of measures that apply only to his or her job responsibilities and only for respondents/customers who have interacted with his/her department or employees. In the past, this kind of reporting would be done after data collection and then disseminated in hardcopy. Portal-based applications can report in real time and disseminate on demand.

Post-Processing in Real-Time

Those directly involved in data collection—both inside and outside of the survey organization—derive considerable value from getting hands on data sooner, sometimes even before data collection is complete. This access might be used to support elements of a responsive design approach, provide for provisional analysis or reporting, or simply as an ongoing quality assurance strategy. We discovered early on that we could create more value for survey data users by building into the data collection phase many of the standard enhancements that normally are made in post processing. Some examples are described below.

Applying Weights

In many instances weights can be designed at the time the sample is pulled, but generally are not applied until they can incorporate the appropriate non-response adjustments. But in the real-time world of many clients there may be value to applying weights early on, especially in ongoing data collections such as tracking studies or when over sampling is used and analysis of unweighted data can be especially misleading. Algorithms can be developed that make it possible to specify weighting schemes and then calculate and apply weights at regular intervals. These intervals might be once a day

or even every time new completed cases are added. To control for broad fluctuations in weighted analyses, a risk especially in the early phases of data collection, the algorithms might require minimum numbers of cases in weighting cells in order to compute these provisional weights. The algorithms might trim the weights within a pre-specified range. At the conclusion of data collection final weights can be computed to improve upon and replace the provisional weights.

Applying Recodes

Researchers often want to ask questions in one format and analyze them in another. For example, we may use scales with varying numbers of points (7, 9, or 11) to capture nuance in respondent attitudes but then recode those data into a fewer number of categories for analysis. If interim data are to be useful, we need to be able to apply these recodes early on and then have them applied to all new data as interviews are completed. The best portals do this.

Creating Composite Variables

Researchers increasingly rely on ever more complex tools to analyze survey data. In MR, competitiveness is often defined by the quality of one's analytic products. For example, a substantial part of MSI's business is the development of statistical models that identify components (groupings of questions) of respondents' attitudes or experiences that contribute to an outcome of interest such as satisfaction with a service or propensity to buy a product. These components in turn become the key variables both in analysis and, in the case of ongoing data collections, monitoring change. This generates a portal requirement to calculate these component variables "on-the-fly," giving clients the ability to monitor change in key measures on an ongoing basis rather than periodically as data collection phases end and component indices are recalculated in post processing.

Merging in Exogenous Data

The addition of respondent or environmental data from other sources is a common technique of data enhancement typically done in post processing. For example, in a B2B study we might want to merge in other data about the respondent's firm from external sources. In a household study we might want to merge in demographic census data about the respondent's town or neighbourhood. As with weights, recodes, and composite variables, the best portals perform these merges on a case-by-case basis as data collection progresses rather in post processing.

4 Looking To The Future

The preceding discussion has focused mostly on what the current generation of Web portals can do or should do today. But as we look to the future, there is significantly more that might be done.

Complex Database Designs

The need to develop analytic datasets from complex, multiple entity data collections is one potential area for development. These data collections might involve relatively straightforward hierarchical models or more complex network designs. Squeezing these designs into standard rectangular data models impinges on analytic flexibility, wastes computing resources, and erects barriers to use. We need easily understood interfaces that allow users to select variables across entities with the underlying links managed automatically and transparently by the portal software.

Consider the example of an educational study that begins with a sample of students and then also collects data on their parents, teachers, school administrators and the school itself. At any point in time, a researcher may want to choose a different unit of analysis or analyze data across entities. He or she might want to look at student characteristics by school characteristics or aggregate up student characteristics based on teacher characteristics. To conduct a meaningful analysis the researcher must comprehend the data model involved, but execution of the linkages across entities needs to be as transparent as possible.

More Powerful Analytics

With some exceptions, the analytic capabilities of the current generation of Web portals are primarily descriptive. It is sometimes easy to export data to more powerful systems and to report on outputs from very complex analyses, but the type of analysis that can be done tends to be basic.

Depending on the audience, there may be an opportunity to deliver more value through more robust analytical capabilities. Correlation, multiple regression, analysis of variance, and simple clustering are among the techniques that might be considered. Another possibility might be to incorporate one of the more powerful Web-based systems described earlier into one's own portal. But once again we make the key point that the value here is integration of these capabilities into the data making process and the ability to do analysis in real-time. Moving data out to other systems adds effort, cost and time, and the connection back to important metadata and paradata can be lost.

Targeted Toolsets

Many if not most surveys are designed to help us understand a specific problem. In the MR world it might be customer satisfaction, the optimal feature set for a new product, or the perceived value of a company brand. In the government and academic sphere it might be educational attainment, labour force participation, or family formation. Likewise, there may be analytical tools that researchers prefer for working with these different kinds of survey datasets. Examples that come to mind include key driver analysis, conjoint, time series analysis, and structural equation modelling.

One can imagine the development of application suites that include the tools commonly used to analyze these different datasets and their incorporation into a Web portal offering. For example, we might present customer satisfaction data with the tools for key driver analysis and trend reporting. We might offer time series analysis tools with labour participation data. Just as the Internet has brought us narrowcasting to tailor information flows to individuals based on their preferences and interests, we might tailor analytical tools to users and the types of data being analyzed.

Integration with Systems Holding Other Data

Macer (2005) recently pointed out that one goal of many Web portal deployments is to create a barrier to clients switching to another research supplier. The thinking goes like this: Once a client's data are stored in such a portal and the client learns to use it, the client is less likely to move to another research supplier for future data collections. Macer expects that clients eventually will resent this and suggests that third party portals such as Knowledge Reporter (www.nunwood.com) may become much more popular with clients because they can aggregate analytic data there from a variety of research

suppliers. If Macer is correct, then portal developers would be wise to build into their portals the capacity to generate compatible files which in the case of Knowledge Reporter, is triple-s.

More immediate is a need increasingly voiced by clients to integrate survey data into their own business intelligence systems, and to do so in real time. Perhaps the most common application is the integration of customer satisfaction data with "hard" customer data stored in the client's internal systems. This in itself is hardly a new application, but what is new is the ability to send survey transactions over the Internet within seconds of a survey's completion.

Respondent Confidentiality and Data Security

No discussion of survey data publication to the Internet is complete without noting the data security challenge that portals pose. While client logins and passwords are *de rigeur* the same cannot always be said about securing sessions with SSL. Nor is there a consensus around what should and should not be stored on a portal. Hard thinking is needed around the issue of what kind of personally identifiable information is permissible on a portal. Should the same standards used for client offline data files apply to portals? Or does their semi-public nature raise the bar?

Once inside a portal a host of additional security concerns arise, often around issues such as who can see what. We increasingly see requirements under which one manager cannot see another's data, the need to restrict what questions an individual can see, and applications that are to be made available to some users but not others. This need to create a tailored experience for individual users is part security and part personalization. It is part of the standard feature set of generic Web portals and something that research companies increasingly must emulate.

5 Summary and Conclusion

Over the last 20 years we have seen a number of truly revolutionary developments in survey research. This paper has described the information value chain as a way to conceptualize those changes and to identify new areas of opportunity. It further argues that the most compelling opportunity at this time to add value to survey data is use of the Internet as a communication and data dissemination platform. By bringing the data user more fully into the data collection process, augmenting data in real-time, delivering faster, and bundling data with the metadata and tools for analysis we are continuing in the CASIC tradition of eliminating steps, reducing cycle time, and economizing on effort.

Of course, this is only today's challenge. As IT evolves there no doubt will be compelling new applications in the value chain. Data collection might be "revolutionized" yet again; analytical tools might evolve that get us to the answers we need more quickly. But at this moment it seems that we produce survey data faster than we can analyze it, and catching up on the analysis is a challenge for all of us. Making it easier should help.

References

R. Baker (1987)
"Information Systems in Survey Research," *Proceedings of the Bureau of the Census Third Annual Research Conference*. Washington, DC: U.S. Bureau of the Census.

R. Blattberg, R. Blazer, and D. Little, eds. (1994)
The Marketing Information Revolution, Boston: Harvard Business School Press.

W. Connett (1998)
"Automated Management of Survey Data," in M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls, and J. M. O'Reilly**,** eds., *Computer-Assisted Survey Information Collection*, New York: Wiley

M. P. Couper (2000)
"Usability Evaluation of Computer-Assisted Survey Instruments," *Social Science Computer Review*, vol 18, 384-396.

M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls, and J. M. O'Reilly, eds. (1998)
Computer-Assisted Survey Information Collection, New York: Wiley.

S. G. Heeringa and R. M. Groves (2004)
"Responsive Design for Household Surveys" Ann Arbor, MI: University of Michigan and Joint Program in Survey Methodology.

A. Jeavons (2002)
"Paradata: Concepts and Applications," presented at Net Effects4, Barcelona.

T. Macer (2005)
"On Tools and Bad Habits: Making MR Technology Fit for the Rigors of Research in the 21st Century," presented at the CASRO 10th Annual Technology Conference, New York.

M. E. Porter (1985)
Competitive Advantage, New York: Free Press.

About the Author

Reginald Baker is Chief Operating Office of Market Strategies, Inc., a full service market research firm specializing in healthcare and pharmaceuticals, electric and gas utilities, financial services, and government and academic research. He is responsible for all of MSI's sampling, interviewing, data processing, analytic, and information technology operations. He has over two decades of experience in all phases of survey research, and has been especially active in the application of new survey technologies. He has authored numerous articles and papers on such subjects as data quality impacts of new data collection technologies, interviewer training for CAPI, CATI system design, and likely future developments in computer-assisted information collection (CASIC). His current interest is the development and evaluation of Web-based survey methods. Prior to joining MSI in 1995, he worked for eleven years at the National Opinion Research Center (NORC) at the University of Chicago where he was Vice President for Research Operations. He can be reached at Market Strategies, 20255 Victor Parkway, Suite 400, Livonia, MI 48105; tel. +01 734 542 7600; email reg_baker@marketstrategies.com

MAXIMISING DATA VALUE. DATA USE & RE-USE
Edited by R. Khan *et al*

Making Existing Data Re-Usable. The Requirements of a Web-enabled Tool

Margaret Ward and Clifford Dive

Abstract

In the past when you wanted to find some information you went along to your local library and either looked for the relevant books yourself, or asked the librarian for help. In today's technological society, this is one of many ways we can locate available information. Today we have access, not only to printed material but also to a vast array of electronic resources, often accessible from our own desks.

In this paper we explore what is required to make data re-usable in our web-enabled world, and also investigate some of the new and exciting developments that have recently become available to the data user of the 21st Century.

Keywords

Metadata; data; accessibility; discovery; analysis; Internet.

1 Introduction

Using the analogy from the days when information was mostly derived from printed material, we would find the information we required from the local library. This provided the means for information to be shared, discovered, analysed, researched and perhaps published in a different form. Today, the principles of the library still apply when looking at the re-use of data, so what does this mean in our web-enabled world?

This paper will look at the issues around accessibility, discovery, metadata, the type of information available, visualisation and analysis and how these are addressed today.

2 Accessibility

The library provides accessibility to books and other resources. It is a place where people go to find what they are looking for – access is often free, and the library can hold vastly more data in one place than any individual could otherwise access. In our modern world a web-enabled data repository can serve the same purpose. Access is physically easy and can be provided free of charge. The user requires no special software on their computer, just a standard web-browser such as Internet Explorer or Mozilla. For sensitive or commercially valuable data, a gatekeeper is required. The human gatekeeper of the old library is replaced by a suitable electronic access control system. Users wishing

to access this protected data are then authenticated, and access granted according to their credentials. These systems can be set up locally so that the owner of a data repository sets their own rules using their own access criteria, or a national system, such as 'Athens' (the national academic registration system) can be used. Control can be set at various levels. For example some users may only have access to an abstract about a particular resource, whereas others may be granted access to the actual raw data, enabling them to conduct their own analysis.

To protect the privacy of individuals when access is being granted to statistical data, we can make use of powerful Statistical Disclosure Control (SDC) procedures. These prevent a sophisticated user combining apparently innocuous data from multiple sources in an attempt to identify individuals.

3 Discovery

None of the above is important or useful unless it is possible to find the required information. A number of strategies can be used in the library. One solution might be to go to the area of the library where appropriate resources are shelved – libraries use standardised classification systems for publications so that the right area can be found easily.

In our web-enabled repository, we need to allow the user to search for, and browse, the available resources. Information therefore needs to be categorised so that the user can browse in the right area, and the user interface needs to present sufficient information to indicate the appropriateness of that information. The use of standardized metadata is therefore essential and this will be discussed later in this paper.

There are times when this approach might not work, and we need to be more focused and use more sophisticated methods. Perhaps a search is appropriate. In today's library environment it is possible to search using an on-line catalogue system. In a web-based system, we can provide search facilities that will allow a simple generic search, or more sophisticated tools to allow the search to be narrowed and focused as required.

Sometimes we know precisely what we are looking for as we have seen a reference to a specific book and an electronic system needs to replicate this. The use of 'bookmarks' provides this functionality. These bookmarks can then be shared by emailing them to interested parties, or embedded into other documents. For example, a report can contain a table, and a reference to the detailed underlying data. If this is an electronic document it is possible to make this reference an active link that can take the user to the original source of that table.

It is possible that the library does not contain the books required by the user. In this case there will be a facility to obtain them from other sources. A web-based system is also capable of accessing resources from multiple sources. For example, Nesstar Ltd. has developed a number of capabilities to support such access. Federated Nesstar Servers provide users with the ability to search for information across a number of Nesstar Servers by using just one search request. The results of that search will be returned and those resources are then available to the user, subject to the relevant authentication procedures. But perhaps one of the most exciting developments is the Nesstar portal. This has been developed as part of the Madiera project (http://www.madiera.net/), whose aim is to provide a web portal to the European Data Archives. Multilingual support is provided through a thesaurus and all datasets are documented according to the DDI (Data Document Initiative) standard. On-line analyses can be performed and the data can also be downloaded in a number of formats. There will also be a Geo-referencing system to enable uses to locate datasets via a map. The portal can harvest multilingual

metadata from a community of Nesstar Servers and provides searching across these pooled resources, whilst the data owner still maintains control and integrity of their data.

Notification

Your library can notify you when certain books become available. A web-enabled system can also provide the facility whereby newly published information is tagged appropriately. This makes it much easier for users to see quickly any new resources that may have been added. It is also possible to notify subscribers when new information, meeting their interest criteria, is published.

4 Describing the data – use of metadata

We might be looking for a 'travel' book about 'Canada'. Travel books are a recognised type of book and a synonym might be 'tourism', or 'exploration'. Similarly, our electronic data can be described by its associated metadata – and this is immensely easier if we have standard forms of describing our information.

Using standardised forms of description for electronic data makes it much easier to locate those resources of interest. The use of metadata standards, e.g. Dublin Core, e-GMS (Government Metadata Standard) and DDI (Data Document Initiative), provide a defined and recognised framework for describing electronic resources. The use of controlled vocabularies, and thesauri, enable the user to locate similar search terms and thus enhance their searching capabilities. This principle may be extended by the use of a multi-lingual thesaurus, to allow searching of data tagged in other languages.

However, to use these standards effectively a tool is needed so that any metadata created is marked up with the appropriate tags. The ideal tool also needs to be able to include definitions of those fields and guidance on how they should be used. Ideally this should be displayed on the input screen so that the person entering the information has easy access to all the information they need.

It is essential that those adopting that particular standard use the metadata fields in the same way, otherwise the information they contain may be misinterpreted. An organisation may only want to use a particular subset of the metadata fields available. Therefore a tool that enables templates containing the chosen fields is particularly useful. These templates can then be shared within an organisation so that all information is marked up in the same way. For the Madiera project, a core set of DDI fields has been selected. These fields must be completed by those joining the Madiera portal, but individual organisations can add to this core set if they wish to include other items of information.

Another important consideration for an organisation is the consistent use, and spelling, of particular words for certain fields. Examples are for key words, topic classifications and the names of geographic areas. The use of a controlled vocabulary for these fields, ideally contained within the template mentioned above, ensures that everyone selects words from a common list thus improving consistency and accuracy of the metadata itself.

5 Support for different document types

The user of the library might find that his research will turn up a variety of document types – official records, books or newspapers. Similarly, web-enabled systems can also support the publishing, and

retrieval, of a range of resource types, including micro-data, tables, MS Word files, 'pdf' files, images and maps.

These resources can also be tagged with the appropriate metadata, thus making them more searchable and therefore accessible.

If an organisation has conducted a survey and reports have also been created. Both the reports and the original micro-data, or tables created from it, can be made available through the same system. This enables the data user to have access not only to the data itself, but also to the associated reports from the same system.

6 Visualisation and analysis

Visualisation

In the library, the book is visualised by removing it from the shelf and reading. The web-enabled system can provide much more than this – the same resources can potentially be viewed in a number of ways.

Published tabular (cube) data can be viewed as a matrix, a graph, or, if appropriately tagged, as a thematic map. However, these are not static views, the data user can decide what to view by using the 'slice and dice', or 'drag and drop' functionality available. This enables any published table, or cross-tabulated variables from a survey, to be re-arranged by moving the table dimensions, thus creating the required layout.

It is also possible to select just those categories of interest for any dimension/variable so that the current view only contains those categories of interest. Dimensions or variables can also be 'moved to layers', whereby only one category of that particular variable/dimension is included in the table at any one time.

Various types of graph including bar charts, pie charts and time series graphs, are also available, and it is also possible to link to various mapping systems. This enables the data displayed in an appropriate table to be visualised on a map.

Although the data publisher will have decided on the default view for a table, once published the user can decide on their own view of that particular table.

Micro-data can be visualised in similar ways, and potentially re-used in numerous ways by the use of cross tabulations using any appropriate variable from the study.

Analysis

Powerful analytical functions are also available, and if more powerful analyses are required it is possible to export, and download, information in a variety of formats, including SPSS and SAS. Some of the data analysis functions available include cross-tabulations, regression analysis, and the ability to recode and compute new variables.

7 Who could make use of such a system?

A web-enabled tool may be used by a variety of organisations. We have already mentioned the Data Archiving community, but there are many other potential users of such technology. Anyone who

wishes to share their data with others could potentially make use of such a system. The Nesstar system has been used by a number of organisations and three examples of such use are detailed below.

Milton Keynes Intelligence Observatory

The Milton Keynes Intelligence Observatory (http://www.mkiobservatory.org.uk) was developed because of the increasing need for stakeholders to work in partnership and hence the importance of sharing data with one another. The Observatory acts as a 'one-stop-shop' for information about Milton Keynes and its surrounding areas and contains data on a wide variety of subjects, including health and the future development of Milton Keynes.

Black Country Observatory

The Black Country Observatory is supported by a number of partners, all of which hold information about the Black Country area. By having a web-enabled tool they have been able to share this information with each other and use it to monitor and evaluate their regeneration policies. For further information please go to: http://www.blackcountryknowledge.co.uk/

Transport for London

Transport for London wanted to create a transport data library on the web that would be useful to the wider transport community. By being web based, it would also benefit users within TfL by consolidating access to the data they hold in a number of disparate systems. This system is now being used by TfL and their stakeholders in the London Transport community.

8 Conclusion

In order for the true value of data to be realised by its re-use, facilities must be provided for resources to be:

- Published - with suitable metadata
- Discovered – by being globally accessible via the Web, with the ability to search and browse
- Visualised - in a variety of ways
- Analysed - using statistical and arithmetic functionality

A variety of systems are available to provide some of these facilities, e.g. Beyond 20/20 (http://www.beyond2020.com/) and Databeacon (http://www.databeacon.com/).

However, the use of a system designed from the ground up using modern Web based technology can provide a very powerful tool for data sharing and re-use across a range of user domains. One example of such a system is Nesstar, which can be seen by visiting: http://www.nesstar.com/.

About the Authors

Margaret Ward works for Nesstar Ltd. and is responsible for managing the technical support function. Margaret can be contacted at Nesstar Ltd., John Tabor House, Wivenhoe Park, University of Essex, Colchester, Essex CO4 3SQ; tel. 01206 872832; email wardm@nesstar.com.

Clifford Dive is the Operations Director for Nesstar Ltd. and can be contacted at Nesstar Ltd., John Tabor House, Wivenhoe Park, University of Essex, Colchester, Essex CO4 3SQ; tel. 01206 874897; email cjdive@nesstar.com.