



RESEARCH

IN OFFICIAL

STATISTICS

I ■ 2001



eurostat

An international journal for research in official statistics

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

Luxembourg: Office for Official Publications of the European Communities, 2001

ISSN 1023-098X

© European Communities, 2001

Printed in France

PRINTED ON WHITE CHLORINE-FREE PAPER

Research in Official Statistics

ROS — An international journal for research in official statistics

ROS — Volume 4 — Number 1 — 2001

Contents

Letter from the editors	5
<i>Photis Nanopoulos, Daniel Defays (Eurostat)</i>	
 Articles	
A decision-theoretic approach to data disclosure problems	7
<i>Mario Trottini</i>	
Uniqueness, urn models and disclosure risk	23
<i>Stephen E. Fienberg and Udi E. Makov</i>	
Structural analysis of the abortion time series reported in the ISTAT summaries	41
<i>Riccardo Bellazzi and Paolo Magni</i>	
Common trends in European school populations	53
<i>Paola Sebastiani and Marco Ramoni</i>	
Computing the posterior distribution of individual-level usual intakes with application to disease models	67
<i>Michael J. Daniels and Alicia L. Carriquiry</i>	
Semi-parametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries	81
<i>Ngiana B. Kandala, Stefan Lang, Stephan Klasen and Ludwig Fahrmeir</i>	
Analysis of aggregated data in survey sampling with application to fertiliser/pesticide usage survey	101
<i>Jaeyong Lee, Christopher Holloman, Alan F. Karr and Ashish P. Sanil</i>	
Bayesian multivariate micro-aggregation under the Hellinger's distance criterion	117
<i>George Kokolakis and Photis Nanopoulos</i>	

Using the national longitudinal survey of youth in the United States to study the birth process: A Bayesian approach <i>Kai Li and Dale J. Poirier</i>	127
Bayesian estimation in a US Census Bureau survey of income recall using respondent- generated intervals <i>S. James Press and Kent H. Marquis</i>	151
On the use of Bayesian networks to analyse survey data <i>Paola Sebastiani and Marco Ramoni</i>	169
On Bayesian record linkage <i>Marco Fortini, Brunero Liseo, Alessandra Nuccitelli and Mauro Scanu</i>	185
Note to authors	

Letter from the editors

One of the objectives of the *Research in Official Statistics* (ROS) journal is to illustrate how official statistics can benefit from research and, through this, promote the use of scientific methods in an area which still has a mainly administrative component.

An opportunity for this is afforded by this issue of the journal which is mostly devoted to Bayesian statistics and its applications in official statistics. The issue is composed of specially selected papers, fulfilling these criteria, contributed and presented at the sixth world meeting of the International Society for Bayesian Analysis. The selection was carried out by a panel from the 234 papers which were presented at the meeting.

Authors of the selected papers were invited to present a more developed version of their papers for the journal. The selected papers were supplemented by two additional ones on the topic of disclosure limitation, because of its special interest to the readers of this journal. All the solicited contributions were then fully reviewed following the standard ROS review process.

The applications presented in this issue cover a broad variety of topics: record linkage, disclosure-risk control and data analysis, to mention just a few. These are all key preoccupations of official statisticians.

The editors would like to thank Professor Stephen E. Fienberg of Carnegie Mellon University and Professor Edward I. George of the University of Texas for their special coordination role in the selection and reviewing process for this issue. We thank you all for your continued interest and support.

Photis Nanopoulos

Daniel Defays

A decision-theoretic approach to data disclosure problems

Mario Trottini

Universitat de Valencia, Spain, and Carnegie Mellon University, United States

Keywords: disclosure limitation, information loss, decision theory, Bayesian theory

Abstract

This paper presents a decision-theoretic approach to data disclosure problems. The approach is innovative because (i) it offers a theoretical framework to develop optimality criteria for the choice of the best form of data release, (ii) it recognises the different perspectives of the statistical agency and of the users of the data in assessing the extent of disclosure and the quality of the users' inference associated with different forms of data release. This leads to new measures of disclosure risk and data utility that take into account not only what the users believe they have learned from the data, but also to what extent their inferences are correct.

1. Introduction

As a part of their activities, most statistical agencies release data sets containing information on individual entities subject to pledges of confidentiality. Nowadays confidentiality is a major legal concern for all statistical agencies as a consequence of the laws governing privacy and confidentiality of statistical data in individual countries and states. In the last 10 years, the amount of statistical data collected has increased enormously and new statistical algorithms and expanding power of computers have increased the danger of disclosure of confidential information. On the other hand, statistical information has become a key element for the actions that both private and public decision-makers have to take, with a corresponding increase in the demand for release of statistical data. Government agencies use statistical data to decide on the allocation of funds and to monitor social programmes, policy analysts use statistical data to inform social decisions, researchers use statistical data to test their theories and to achieve a deeper understanding of the phenomena under study. Ideally a statistical agency should provide maximum information to the users preserving the privacy of the individual entities represented in the data set. The subfield of statistics concerned with such a problem is usually referred to as statistical data protection, statistical confidentiality or statistical disclosure limitation.

In this paper, we present a decision-theoretic approach to data disclosure problems. In our disclosure scenario it is assumed that the set of users can be partitioned in two groups. Those who want to use the released data to perform statistical studies or for research purposes, and those who want to use the released data to disclose confidential information about the data providers. We refer to the first group as society and to the second group as the intruder. More

formally, we assume that society is interested in an unknown quantity Θ_{SOC} while the intruder is interested in an unknown quantity Θ_{INT} and they try to infer their target values based on the information contained in the released data. We assume the availability of a specified set of alternative forms of data release and the agency must decide which is best to release, taking into account the extent of disclosure and the data utility associated with each form of data release. The extent of disclosure measures the extent to which the intruder's inference about Θ_{INT} , based on the released data, can harm the data providers and/or the statistical agency. This is balanced by the data utility that measures the extent to which the released data is useful to society. We suggest choosing as the optimal form of data release the one that maximises the data utility among those releases whose extent of disclosure falls below a fixed threshold.

Our measures of the extent of disclosure and data utility are generalisations of measures proposed by Lambert (1993) for re-identification problems. The underlying idea is that in assessing the extent of disclosure and the data utility, we need to take into account not only what the users believe they learn from the released data but also what they actually learn. Most current measures take into account only the first component (see for example, Duncan and Lambert (1986), Duncan and Pearson (1991), Fienberg et al. (1997)). In this paper we argue that it is difficult to measure harm (and thus disclosure) and the usefulness of data if we ignore the correctness of the users' inferences, since correct and incorrect inferences may have different consequences.

Sections 2 and 3 contain a brief literature review of the measures of disclosure and data utility currently used in statistical confidentiality. In particular, in Section 2 we review in some detail the works of Duncan and Lambert (1986) and Lambert (1993) that inspired the main ideas of this work. Section 4 addresses the problem of the role of the statistical agency in assessing the disclosure risk and data utility. Section 5 sets up the basic assumptions and notation. Sections 6 and 7 introduce new definitions of disclosure risk and data utility and describe a general framework for the representation of the trade-off of gains versus risk. Section 8 presents an optimality criterion for the choice of the best form of data release, and Section 9 contains concluding remarks.

2. Current measures of the extent of disclosure

Most of the approaches currently used in statistical confidentiality to measure the extent of disclosure equate disclosure with what the intruder believes has been disclosed. The work of Duncan and Lambert (1986) provides a unifying framework for such methods. The approach is as simple as it is powerful. It assumes that the intruder prior beliefs about Θ_{INT} before and after the release of the data can be expressed in terms of probability distributions that the authors refers to as prior and posterior (intruder's) predictive distributions respectively. Different measures of disclosure are obtained by applying an uncertainty function to the prior

and posterior distributions of the intruder. An uncertainty function is a non-negative measurable mapping from the space of all possible distributions to the set of non-negative real numbers. Each distribution is associated with a non-negative number. The larger the number the more the uncertainty about the value of the random variable with that distribution. In particular if $L_{\Theta_{INT}}^{(I)}$ is the loss function for the intruder's decision problem 'identify Θ_{INT} ', the uncertainty function is defined as the risk associated with the optimal estimate of Θ_{INT} with respect to $L_{\Theta_{INT}}^{(I)}$. Denote by $p^{(I)}$ the probability distribution function that formalises the intruder's beliefs about Θ_{INT} . The intruder's uncertainty about the true value of Θ_{INT} is then given by:

$$U_{\Theta_{INT}}^{(I)}(p^{(I)}) = \operatorname{argmin}_a \int L_{\Theta_{INT}}^{(I)}(a, \mathcal{G}_{INT}) p^{(I)}(\mathcal{G}_{INT}) d\mathcal{G}_{INT}$$

Based on the idea that the larger is the intruder's uncertainty after the release of the data, the smaller is the risk of disclosure, the different measures of disclosures are defined as decreasing functions of $U_{\Theta_{INT}}^{(I)}$ (posterior). The disclosure rule is then to release the data if and only if $U_{\Theta_{INT}}^{(I)}$ (posterior) is bigger than a fixed threshold.

Duncan and Lambert (1986) show that different measures of the extent of disclosure currently used by statistical agencies can be obtained as special cases of this general framework for suitable choices of the loss function $L_{\Theta_{INT}}^{(I)}$. For example, suppose that the released data consists of a cross-classification with k categories. Assume that the intruder's target is to disclose to which of the k categories a particular respondent, say A, belongs and that the intruder's loss, when he incorrectly specifies that A belongs to category i , is a decreasing function of the posterior probability that A belong to that category. Duncan and Lambert's disclosure rule, for this case, provides a basis for the ad-hoc rules for bounding tabular relative frequencies away from zero and away from one, discussed by the Subcommittee on Disclosure Avoidance Techniques (1978) and currently used by many statistical agencies. The defining feature of Duncan and Lambert's approach is that the extent of disclosure only depends on the intruder's uncertainty. Disclosure takes place if the released data make the intruder confident about his inference but the method makes no distinction between correct and incorrect inferences. The underlying idea is that correct disclosure and incorrect disclosure are both dangerous and should be avoided.

Lambert (1993) argues that it is difficult to measure harm taking into account only the intruder's uncertainty since the consequences of correct and incorrect disclosure are usually different. In particular for re-identification problems, Lambert distinguishes between the risk of perceived identification (the maximum of the intruder's probability that one of the released records in the source file is the target's) which represents what the intruder believes has been disclosed, and the risk of true identification (the percentage of records correctly identified by the intruder) which represents what the intruder actually has disclosed. This more general definition of disclosure recognises the complexity and the variety of ways in which disclosure can occur. If an agency's goal is to prevent only correct inferences then the risk of true identification is the appropriate measure of disclosure. If, instead, the agency only wants to

prevent an intruder from believing that he has disclosed confidential information then it should use the risk of perceived identification. Finally if it wishes to prevent both correct inferences and intruder's perceived disclosure, it should take into account both the risk of true identification and the risk of perceived identification. Lambert's approach is flexible enough to provide a suitable measure of disclosure for all of these situations. Unfortunately the approach is problem specific and applies only to re-identification problems. The measure of global risk that we present in Section 6 generalises this approach and provides a general framework to measure the extent of disclosure for an arbitrary disclosure limitation problem.

3. Current measures of data utility and trade-off of gain versus risk

Considerable research has been devoted to the assessment of the extent of disclosure, but little has been done to measure the impact of statistical disclosure techniques on statistical analyses. Most approaches proposed to date have been problem specific and do not provide a general framework to measure data utility (e.g., see Mateo-Sanz and Domingo-Ferrer (1999), Baeyens and Defays (1999), de Waal and Willenborg (1998) and Hurkens and Tiourine (1998)). Willenborg and de Waal (2001) and Trottni (2001) represent efforts to build such a framework. Willenborg and de Waal's measure of data utility compares the entropy of the original data with the entropy of the released data. The method is very general and simple to implement; however, it ignores the final use of the data and it doesn't describe the society's behaviour when data are released. On the other hand, the measures of data utility in Trottni (2001) depend only on society's posterior uncertainty about the true value of its target but do not take into account how accurate is society's inference. The lack of suitable methods to measure the data utility has resulted in a lack of criteria to compare alternative forms of data release that are able to take into account the trade-off between disclosure risk and data utility.

Duncan and Keller-McNulty (2001) suggest using the uncertainty measures proposed in Duncan and Lambert (1986) for the extent of disclosure and the MSE of the estimator of society's target for data utility. The optimal form of data release then maximises the data utility among those possible releases whose extent of disclosure falls below a fixed threshold. Their work emphasises the trade-off of gain versus risk rather than the development of an appropriate measure of data utility. Trottni (2001) modifies this approach by using the measures of disclosure discussed in Duncan and Lambert (1986) for data utility instead of MSE. But in neither approach does the optimality criteria distinguish between correct and incorrect users' (intruder's and society's) inferences. In Sections 7 and 8, we present measures of data utility and an optimality criterion that allow for this distinction.

4. Correct and incorrect inferences: The role of the statistical agency

In the literature on disclosure limitation, the role of the statistical agency in assessing the risk of disclosure and the data utility has undergone very little exploration. Most authors presume that the agency only tries to reproduce what would be the intruder's and the society's behaviour when the data are released (see for example Duncan and Lambert (1986)). In the few cases in which the agency plays a role in the assessment of the risk of disclosure or data utility (as in the re-identification problem), this role is not explicitly recognised and as a consequence it is not formalised. We believe that the agency's perspective should be also a component of a disclosure limitation problem since it allows distinguishing between correct and incorrect inferences of the data users (intruder and society). Consider the following example.

Example 1. Suppose that two data sets D_1 and D_2 are considered for release. Suppose that the intruder's posterior distributions for Θ_{INT} given D_1 and D_2 are $N(-5,1)$ and $N(5,1)$ respectively while the society's posterior distributions for Θ_{SOC} given D_1 and D_2 are $N(-7,2)$ and $N(7,2)$ (here $N(m, v)$ denotes a normal distribution with mean m and variance v). Suppose also that the agency's posterior distributions for Θ_{INT} and Θ_{SOC} given the original data are $N(4.9,0.001)$ and $N(-6.9,0.001)$ respectively and that intruder, society and agency use a quadratic loss function. Under quadratic error loss, the optimal action is the posterior mean and the uncertainty (i.e. the expected loss associated with the optimal action) is the posterior variance. Thus the intruder's optimal estimates of Θ_{INT} when D_1 and D_2 are released are -5 and 5, respectively, while the intruder's uncertainty in both cases is 1. Similarly, society's optimal estimates of Θ_{SOC} when D_1 and D_2 are released are -7 and 7, respectively, and society's uncertainty in both cases is 2. Based on the measures of disclosure in Duncan and Lambert (1986) as adapted by Trottni (2001), the agency should flip a coin to decide which data set is best to release, since in either case the intruder's uncertainty is 1 and the society's uncertainty is 2, and D_1 and D_2 are perfectly equivalent. However the agency posterior distribution for Θ_{INT} and Θ_{SOC} are $N(4.9,0.001)$ and $N(-6.9,0.001)$.

In this example, the agency is very confident that the true value of Θ_{INT} is approximately 4.9 (the 95 % posterior credibility interval for Θ_{INT} is [4.84,4.96]) and the true value of Θ_{SOC} is -6.9 (the 95 % posterior credibility interval for Θ_{SOC} is [-6.96, -6.84]). If D_2 is released the intruder's optimal estimate of Θ_{INT} is 5. This is very close to what the agency's believes to be the true value of Θ_{INT} while the society's optimal estimate of Θ_{SOC} , 7, is very poor from the agency's point of view. If D_1 is released, instead, the intruder's optimal estimate of Θ_{INT} , -5, is very inaccurate from the agency's point of view while the society's optimal estimate of Θ_{SOC} ,

-7, is very close to what the agency believes to be the true value of Θ_{SOC} . Thus if we take into account not only the intruder's and society's perspectives, but also the agency's knowledge of the intruder's and the society's targets Θ_{INT} and Θ_{SOC} , the two forms of data release, D_1 and D_2 , are no longer equivalent and D_1 should be intuitively released.

In Sections 6 and 7 we present new measures of disclosure and data utility that formalise the role of the agency in assessing disclosure risk and data utility. In Section 8, based on these new measures of disclosure and data utility, we return to this example, and show that, in accord with intuition, the release of D_1 should be preferred to the release of D_2 . But first, in the next section we introduce some basic notation and assumptions that characterise the new measures.

5. Notations and general assumptions

As a result of its activities, a statistical agency produces a data set D_0 . In order to reduce the risk of disclosure of confidential information, the agency modifies the original data using different disclosure limitation techniques (a review of alternative disclosure limitation techniques can be found in Willenborg and de Waal (2001)). These produce a class \mathbf{D} of alternative forms of data release. We denote by D_R the generic element in \mathbf{D} . The goal of the agency is to choose the best form of data release in \mathbf{D} . The definition of an optimality criterion requires some assumptions about the behaviour of the intruder, society, and statistical agencies, how they formalise their prior information about the target values, how they update this prior information, and how they use the released data.

To make the problem meaningful, we assume that both Θ_{INT} and Θ_{SOC} are somehow related to the original data D_0 . In particular, we assume that, prior to observing D_0 , the statistical agency believes that D_0 is a realisation of a random variable V_A whose distribution P_A belongs to a parametric family \mathbf{PA} with parameter Ψ_A and parameter space Ω_A . The intruder and society, prior to observing D_0 , believe that D_0 is a realization of random variables V_{INT} and V_{SOC} whose distributions P_{INT} and P_{SOC} belong to parametric families \mathbf{PINT} and \mathbf{PSOC} with parameters Ψ_{INT} and Ψ_{SOC} and parameter spaces Ω_{INT} and Ω_{SOC} respectively. The conditional distributions of V_A , V_{INT} , and V_{SOC} given Θ_{INT} , and the conditional distributions of V_A , V_{INT} , and V_{SOC} given Θ_{SOC} , formalize how the data D_0 are related to the intruder's and society's targets, Θ_{INT} and Θ_{SOC} . We also assume that the intruder's and society's prior beliefs about Ψ_{INT} , Θ_{INT} , and Ψ_{SOC} , Θ_{SOC} , can be adequately expressed by probability distributions, $\pi_{\Psi_{INT}}(\cdot)$, $\pi_{\Theta_{INT}}^{(I)}(\cdot)$, $\pi_{\Psi_{SOC}}(\cdot)$, $\pi_{\Theta_{SOC}}^{(S)}(\cdot)$, that we refer to as *intruder's and society's prior distributions* for Ψ_{INT} , Θ_{INT} , and Ψ_{SOC} , Θ_{SOC} respectively. Similarly we assume that the agency's prior beliefs

about Ψ_A , Θ_{INT} , and Θ_{SOC} can be adequately expressed by probability distributions, $\pi_{\Psi_A}(\cdot)$, $\pi_{\Theta_{INT}}^{(A)}(\cdot)$, $\pi_{\Theta_{SOC}}^{(A)}(\cdot)$ that we refer to as *agency's prior distributions* for Ψ_A , Θ_{INT} , and Θ_{SOC} respectively. The conditional distribution of V_{INT} given Ψ_{INT} , and V_{SOC} given Ψ_{SOC} , the prior distributions for Ψ_{INT} and Ψ_{SOC} , and the particular disclosure limitation technique used induce, for each form of data release, D_R , conditional distributions of D_R given Θ_{INT} and D_R given Θ_{SOC} that formalize how the released data set D_R is related to the intruder's and society's targets Θ_{INT} and Θ_{SOC} . We denote by $\pi_{\Theta_{INT}}^{(I)}(\cdot|D_R)$ the conditional distribution of Θ_{INT} given D_R and by $\pi_{\Theta_{SOC}}^{(S)}(\cdot|D_R)$ the conditional distribution of Θ_{SOC} given D_R . These are the *intruder's and society's* posterior distributions for Θ_{INT} and Θ_{SOC} given D_R and they express the intruder's and society's beliefs about their targets, after the data D_R have been released. Similarly, we denote by $\pi_{\Theta_{INT}}^{(A)}(\cdot|D_0)$ and $\pi_{\Theta_{SOC}}^{(A)}(\cdot|D_0)$ the conditional distributions of Θ_{INT} and Θ_{SOC} given D_0 . These are the agency's posterior distributions for Θ_{INT} and Θ_{SOC} given D_0 and they express the agency's beliefs about the target values Θ_{INT} and Θ_{SOC} , after the data D_0 have been observed. Note that the intruder's and society's posterior distributions are usually different from the agency's posterior distributions. This is not just because the agency's priors for Θ_{INT} and Θ_{SOC} might differ from the intruder's and society's priors, or because the agency's model might differ from the intruder's and society's models, but also because the statistical agency possesses the original data D_0 . Therefore the agency updates its beliefs about Θ_{INT} and Θ_{SOC} using the original data D_0 instead of the released data D_R .

We also assume the following:

Assumption 1: The statistical agency knows the intruder's and society's loss functions, their prior distributions for the targets values Θ_{INT} , Θ_{SOC} as well as their uncertainties about the model generating the original data D_0 .

Assumption 1 is not necessarily realistic; however, we can easily relax it to fit more realistic scenarios. The statistical office can use classes of prior distributions and classes of loss functions to describe the intruder's and society's prior uncertainties about their targets, and the loss that the intruder and society are willing to pay for a generic estimate of their target values. Similarly, we can use classes of distributions to describe the intruder's and society's uncertainties about the mechanism producing the data D_0 .

Assumption 2: The statistical office releases complete information about the mechanism that produces the released data set.

Many statistical agencies do not release complete information about the disclosure limitation techniques they use, e.g., the parameter values in the concentration rule for cell suppression (see Duncan et al., 1993). Our position is that statistical data are a public good (see Fienberg, 2000), and thus the statistical agency should release as much information as possible about the

mechanism generating the released data set. The more information that is available the easier it is for an observer to predict the behaviour of the intruder and society and therefore to make correct inferences about the disclosure risk and the data utility associated with the released data.

Assumption 3: Both the intruder and society act rationally, according to the expected loss principle, i.e. in estimating their targets they try to minimise the posterior risk.

The approach we propose is normative. We describe what the intruder and society should do rather than what they actually do in practice. From the statistical agency's point of view, a descriptive approach might seem more appropriate. A descriptive approach is much more difficult to implement, however, and when assumption 2 is satisfied there should not be a big difference between the two approaches.

Assumption 4: The probability that the intruder takes actions is a decreasing function of the intruder's uncertainty.

Assumption 4 implicitly assumes that the intruder has to pay a penalty when he claims that confidential information has been disclosed, but actually no disclosure has taken place. This assumption is realistic in many problems (e.g., when legal systems allow for redress of harm resulting in the misuse of public information), although not in general.

6. Representing disclosure risk

The choice of the best form of data release requires a notion of disclosure risk and data utility. In a broad sense the disclosure risk associated with the release of a data set D_R is a measure of the extent to which the release of D_R makes it possible for the intruder to create harm. For example, the intruder's goal could be to obtain information about particular individuals, to discredit the agency, or simply to show his own cleverness. It seems reasonable for the statistical agency to act 'as if' all these goals actually co-exist. As a result we make the following assumption:

Assumption 5: The intruder can create harm in two different ways:

1. **disclosure harm:** the intruder discloses confidential information about the providers of the data, i.e. the intruder's inference is correct;
2. **discredit harm:** the intruder discredits the statistical agency or the data providers, claiming that confidential information has been disclosed.

The notions of disclosure harm and discredit harm generalise the definitions of risk of true identification and risk of perceived identification introduced by Lambert (1993) for re-identification problems. Disclosure harm takes place when the intruder's inference about Θ_{INT} is correct, (for example when the intruder correctly identifies a respondent whose record is in a released microdata set, or gives a very good approximation of the value of a sensitive attribute of a respondent). In the re-identification context considered by Lambert, the agency knows the true value of the intruder's target and therefore it makes sense to talk of true and false disclosure (since the intruder is either right or wrong). In a more general setting, the agency

can only estimate the true value of the intruder's target, and the extent to which the intruder's inference is correct does not depend only on his inference but also on the agency's inference about Θ_{INT} . Discredit harm is, instead, a measure of how confident the intruder is about his inference and can occur even when the intruder's inference is completely incorrect but the intruder believes that his inferences are precise and acts accordingly. The two types of harm can occur in different ways and we need to distinguish among them. We introduce the following definitions:

Definition 1: Let $U_{\Theta_{INT}}^{(I)}(p^{(I)})$ denote the intruder's uncertainty about Θ_{INT} when his distribution over Θ_{INT} is $p^{(I)}(\cdot)$. We define the risk of discredit harm (RDH) when the intruder's distribution over Θ_{INT} is $p^{(I)}(\cdot)$ as:

$$\text{RDH}(p^{(I)}) = -U_{\Theta_{INT}}^{(I)}(p^{(I)}) \quad (2)$$

Definition 1 rephrases and specialises assumption 4. The risk of discredit harm is a decreasing function of the intruder's posterior uncertainty about Θ_{INT} . Here we use the function 'minus the uncertainty', but other choices, of course, are possible. RDH is minus the (intruder's) posterior knowledge measure of disclosure proposed by Duncan and Lambert (1986) and thus cannot be positive. $\text{RDH} = 0$ if and only if the intruder's has no uncertainty about the true value of the target Θ_{INT} .

Definition 2: Let G_A and G_{INT} be two probability distributions, and let $D(G_A, G_{INT})$ be a measure of divergence between G_A and G_{INT} . $D(G_A, G_{INT})$ is a measure of how well G_{INT} approximates G_A . Also let $U_{\Theta_{INT}}^{(A)}(p^{(A)})$ be the agency's uncertainty about Θ_{INT} when its distribution over Θ_{INT} is $p^{(A)}(\cdot)$. We define the intruder's estimated knowledge (IEK) from the agency's point of view when the intruder's distribution over Θ_{INT} is $p^{(I)}(\cdot)$ as:

$$\text{IEK}(p^{(I)}) = -[D(\pi_{\Theta_{INT}}^{(A)}(\cdot | D_0), p^{(I)}) + U_{\Theta_{INT}}^{(A)}(\pi_{\Theta_{INT}}^{(A)}(\cdot | D_0))] \quad (3)$$

In particular we define the intruder's prior estimated knowledge, $\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\cdot))$, and the intruder's posterior estimated knowledge, $\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\cdot | D_R))$.

Definition 2, says that, from the agency's point of view, the intruder's knowledge about Θ_{INT} when his distribution over Θ_{INT} is $p^{(I)}$, is extensive (i.e. the risk of disclosure harm is high) only if the agency's uncertainty about Θ_{INT} based on the original data D_0 is very small and the intruder's distribution over Θ_{INT} , $p^{(I)}$, does not differ too much from the agency's posterior distribution based on the original data, D_0 .

Note that

$$\text{IEK} \leq -U_{\Theta_{INT}}^{(A)}(\pi_{\Theta_{INT}}^{(A)}(\cdot | D_0)) \leq 0 \quad (4)$$

and $\text{IEK} = 0$ if and only if the agency's has no uncertainty about the true value of the target Θ_{INT} and the agency's and intruder's distributions over Θ_{INT} are the same, as measured by the divergence $D(\cdot, \cdot)$.

A reasonable measure of disclosure associated with the release of a data set D_R should take into account both the risk of disclosure harm and the risk of discredit harm. We propose the following definition of global risk.

Definition 3: We define the global risk associated with the release of a data set D_R as the vector:

$$\text{G.RISK}(D_R) = [\text{IEK}(\pi_{\Theta_{INT}}^{(I)}(\cdot | D_R)), \text{RDH}(\pi_{\Theta_{INT}}^{(I)}(\cdot | D_R))] \quad (5)$$

Using definition 3, we can represent the risk associated with a form of data release D_R as a point $p(D_R)$ in a Cartesian plane, with coordinates $\text{G.RISK}(D_R)$, as shown in Figure 1 – where t_{INT} and t_A denote threshold values for the maximum tolerable risk of discredit harm and intruder's estimated knowledge respectively.

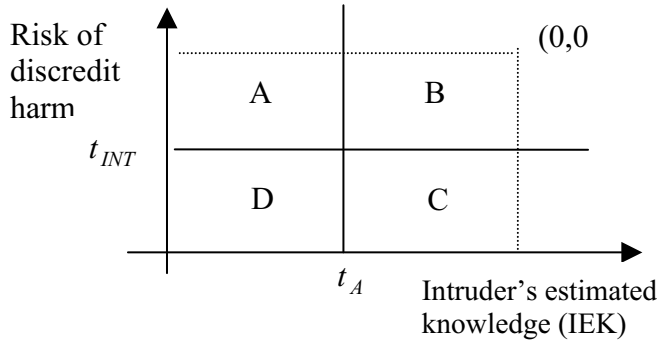


Figure 1: Global risk

We assume that, prior to the release of the data, both the risk of discredit harm and the intruder's estimated knowledge are below the corresponding thresholds. Then points in A correspond to data sets for which the intruder's inference is imprecise from the agency's point of view, however, the intruder is confident about his inference. In this case, from the agency's point of view there is no violation of confidentiality but the intruder is likely to take some actions and act 'as if' confidentiality has been violated. Thus we have discredit harm but not disclosure harm. Points in B correspond to data sets for which the intruder's inference is very precise both from the agency's and intruder's points of view. In this case, from the agency's

point of view the intruder is very likely to correctly identify confidential information and act upon it. We have both disclosure harm and discredit harm. Points in C correspond to data sets for which the intruder's inference is correct (from the statistical office point of view), but the intruder is not very likely to take any action since his uncertainty about the target value is very high. Thus we have disclosure harm but not discredit harm. Finally, points in D correspond to data sets for which there is no violation of confidentiality and the intruder is not likely to take any action, since his uncertainty about the true value of his target is very high. Data sets corresponding to points in D are the safest; those corresponding to points in B are the most dangerous. In particular the point (0,0) in Figure 1, corresponds to a data set whose release allows the intruder to disclose the true value of his target with probability one.

Note that, as in Lambert (1993), correct and incorrect inferences can be distinguished if desired but they need not be. An appropriate choice of the threshold values, t_{INT} and t_A , in principle allows the agency to prevent only correct inference (i.e. disclosure harm), or only discredit harm or both. In particular, if the goal is only to prevent the correct inferences, then t_{INT} should be set equal to zero and t_A set at an appropriate value strictly less than zero (the smaller t_A the bigger the protection from disclosure harm). If, instead, the goal is only to prevent the intruder from believing that he has disclosed confidential information, then t_A should be set equal to zero and t_{INT} set at an appropriate value strictly less than zero (the smaller t_{INT} the bigger the protection from discredit harm).

7. Representing the data utility

The measure of global risk, described in the previous section, has a natural extension as a measure of data utility. In a broad sense the data utility associated with the release of a data set D_R is a measure of the extent to which the release of D_R makes it possible for society to make accurate inferences about its target Θ_{SOC} . As for the extent of disclosure, we need to distinguish between the agency and the users' perspectives; between correct and incorrect (society's) inferences. We introduce the following definitions.

Definition 6: Let $U_{\Theta_{SOC}}^{(S)}(p^{(S)})$ be society's uncertainty about Θ_{SOC} when its distribution over Θ_{SOC} is $p^{(S)}(\cdot)$. We define the perceived data utility (PDU) of a data user as society's posterior knowledge (c.f., Trottini, 2001):

$$PDU(p^{(S)}) = -U_{\Theta_{SOC}}^{(S)}(p^{(S)}) \quad (6)$$

Note that $PDU \leq 0$ and $PDU=0$ if and only if society has no uncertainty about the true value of the target, Θ_{SOC} .

Definition 7: Let $D(.,.)$ be a measure of divergence, as in definition 2. Also let $U_{\Theta_{SOC}}^{(A)}(p^{(A)})$ be the agency's uncertainty about Θ_{SOC} when its distribution over Θ_{SOC} is $p^{(A)}(.)$. We define the society's estimated knowledge (SEK) from the agency's perspective, when society's distribution over Θ_{SOC} is $p^{(S)}(.)$, as:

$$SEK(p^{(S)}) = -[D(\pi_{\Theta_{SOC}}^{(A)}(\cdot|D_0), p^{(S)}) + U_{\Theta_{SOC}}^{(A)}(\pi_{\Theta_{SOC}}^{(A)}(\cdot|D_0))] \quad (7)$$

In particular we use expression (7) for society's prior estimated knowledge, $SEK(\pi_{\Theta_{SOC}}^{(S)}(.))$, and society's posterior estimated knowledge, $SEK(\pi_{\Theta_{SOC}}^{(S)}(\cdot|D_R))$. Note that $SEK \leq -U_{\Theta_{SOC}}^{(A)}(\pi_{\Theta_{SOC}}^{(A)}(\cdot|D_0))$ and $SEK=0$ if and only if the agency has no uncertainty about the true value of Θ_{SOC} and the agency's and society's distributions over Θ_{SOC} are the same, as measured by the divergence $D(.,.)$.

Perceived data utility (similar to the risk of discredit harm) measures how confident society is about its inferences. Society's estimated knowledge (similar to the intruder's estimated knowledge) measures how good society's inference is from the agency's perspective. A reasonable measure of data utility should take into account both society's perceived data utility and society's estimated knowledge. We propose the following definition of global utility:

Definition 8: We define the global data utility associated with the release of a data set D_R as:

$$G.UTILITY(D_R) = [SEK(\pi_{\Theta_{SOC}}^{(S)}(\cdot|D_R)), PDU(\pi_{\Theta_{SOC}}^{(S)}(\cdot|D_R))] \quad (8)$$

Using definition 8, we can represent the data utility associated to each form of data release D_R as a point $q(D_R)$ in a Cartesian plane, with coordinates $G.UTILITY(D_R)$, as shown in Figure 2.

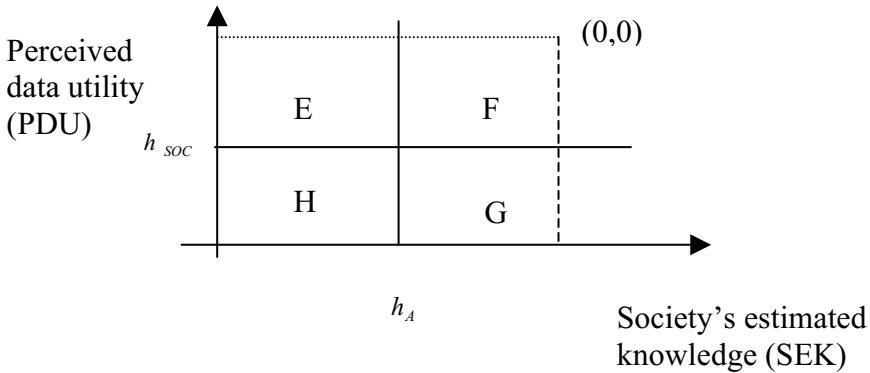


Figure 2: Global utility

Using threshold values h_{SOC} and h_A to discriminate between high and low values of the perceived data utility and society's estimated knowledge, we can divide the plane into four sectors. Points in E correspond to data sets for which society's inference is very imprecise (from the agency's point of view), however society is very confident about its inference. Points in F correspond to data sets for which society's inference is very precise both from the society's and the agency's point of view. In particular the point (0,0) represents a form of data release with maximum global utility that leads society to infer the correct value of its target with probability one. Finally, points in G correspond to data sets for which society's inference is correct (from the statistical office point of view), but society is very uncertain about the true value of its target Θ_{SOC} while points in H correspond to data sets for which society's inference is inconclusive both from the agency's and society's point of view.

8. An optimality criterion

Using the measures of global risk and global utility introduced in Sections 6 and 7, we propose the following optimality criterion for the choice of the best form of data release:

Optimality criterion: Consider a given disclosure scenario (i.e. an original data set D_0 , the intruder's target, the society's target, their prior distributions, their loss functions, the agency's prior distributions, the agency's loss functions, etc.). Let t_{INT} and t_A be threshold values for the risk of discredit harm and for the intruder's estimated knowledge respectively and let \mathbf{D} be the class of the alternative forms of data release. Denote by D_{1A} the subset of \mathbf{D} containing all data set in \mathbf{D} that have risk of discredit harm and intruder's estimated knowledge (with respect to the given disclosure scenario) below the corresponding thresholds (i.e. D_{1A} consists of all data sets in \mathbf{D} whose global risk belong to the region D in Figure 1). If D_{1A} is empty then do not release any data set. If D_{1A} is not empty then release the data set in D_{1A} whose global utility (with respect to the given disclosure scenario) minimises the Euclidean distance from the point (0,0).

The idea underlying this criterion is very simple. Among all safe data sets, i.e. those whose global risk is in the region D in Figure 1, the optimal one is the one whose global utility is closest to the point of maximum global utility (0,0). If a safe data set does not exist, no data set is released.

Example 1 reconsidered. Suppose that an agency uses the Kullback-Leibler information statistic as its measure of divergence. Then the global risk of D_1 is $G.RISK(D_1) = (-51.96, -1)$ the global risk of D_2 is $G.RISK(D_2) = (-2.96, -1)$, the global utility of D_1 is $G.UTILITY(D_1) = (-3.30, -2)$ and the global utility of D_2 is $G.UTILITY(D_2) = (-51.60, -2)$. The data sets D_1 and D_2 have the same risk of discredit harm (-1). This means that if the agency releases either D_1 or D_2 , the intruder is equally likely to discredit the agency claiming that

confidential information has been disclosed. Nonetheless, the intruder's estimated knowledge when D_1 is released is much smaller than the intruder's estimated knowledge when D_2 is released, i.e. from the agency's point of view the release of D_2 leads to a much better intruder's inference than the release of D_1 . Similarly, D_1 and D_2 have the same perceived data utility (-2), i.e., if either D_1 or D_2 is released, society's uncertainty about the true value of the target Θ_{SOC} is the same. However, society's estimated knowledge when D_1 is released is much higher than the society's estimated knowledge when D_2 is released. This means that, from the agency's point of view, society's inference when D_1 is released is much better than society's inference when D_2 is released. Using threshold values $t_{INT} = -0.8$ and $t_A = -40$ there is only one safe data set, D_1 , and therefore, according to the optimality criterion, the agency should release D_1 . As we expected, D_1 is always the optimal form of data release, no matter which threshold values we use except for the case in which no safe data set exists and therefore no data set should be released.

9. Conclusions

In this paper we present a decision-theoretic framework to measure the extent of disclosure and the data utility. The defining feature of our approach is that disclosure risk and data utility do not depend only on the released data and on the users' (intruder's and society's) behaviour but also on the agency's knowledge of the intruder's and society's targets. It is the agency's knowledge of these targets that define the extent to which the users' inferences are correct. The measure of disclosure that we present in Section 6 (the global risk) is a vector with two components. The first component (the risk of discredit harm) is a measure of what the intruder believes has been disclosed; the second component (the intruder's estimated knowledge) measures what the intruder actually has disclosed (taking into account the agency's knowledge of the intruder's target). The appealing feature of this representation of the extent of disclosure is that it is general enough to represent the variety of ways in which people (data providers and agencies) think about disclosure. As we showed in Section 6, the global risk is compatible with different definitions of disclosure. With a suitable choice of the threshold values for the risk of discredit harm and the intruder's expected knowledge we can prevent only correct inferences (disclosure harm), or we can prevent only the intruder from believing he has disclosed confidential information (discredit harm) or both. Our measure of data utility is a natural extension of the measure of disclosure. The global utility is a vector whose components express respectively what society believes it has learned about its target and what society actually has learned about its target from the released data. When we reconsidered example 1, we showed that optimal forms of data release using our criterion differ substantially from those that agencies traditionally adopt.

10. Acknowledgements

Preparation of this paper was supported in part by a Marie Curie Fellowship of the European Community programme ‘Improving the human research potential’ under contract number HPMFCT-2000-00463, with Host Institution the Department of Statistics and Operative Research of the University of Valencia; in part by the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences and a subcontract to Carnegie Mellon University. The content of the paper reflects the author’s personal opinion. The European Commission and the National Institute of Statistical Sciences are not responsible for any views or results expressed. I would like to thank M. J. Bayarri and S. E. Fienberg for their comments and support during the preparation of the paper.

11. References

- [1] Baeyens, Y. and Defays, D. (1999), ‘Estimation of variance loss following from micro-aggregation by individual ranking method’, *Statistical Data Protection, Proceedings of the Conference*, Lisbon, Portugal, Eurostat, pp. 101–108.
- [2] de Waal, A. G. and Willenborg, L. C. R.J. (1998), ‘Optimal local suppression in microdata’, *Journal of Official Statistics*, Vol. 14, pp. 421–435.
- [3] DeGroot, M. H. (1962), ‘Uncertainty, information and sequential experiments’, *Annals of Mathematical Statistics*, Vol. 3, pp. 404–419.
- [4] Duncan, G. T. and Keller-McNulty, S. (2001), ‘Risk of statistical confidentiality disclosure: A comparison of masked and synthetic data release’, *Research in Official Statistics*, Vol. 3, in press.
- [5] Duncan, G. T. and Lambert, D. (1986), ‘Disclosure-limited data dissemination’, *Journal of the American Statistical Association*, Vol. 81, pp. 10–28.
- [6] Duncan, G. T. and Pearson, R. B. (1991), ‘Enhancing access to microdata while protecting confidentiality: Prospects for the future’ (with discussion), *Statistical Science*, Vol. 6, pp. 219–239.
- [7] Duncan, G. T., Jabine, T. B. and De Wolf, V. A. (eds) (1993), *Private lives and public policies: Confidentiality and accessibility of government statistics*, National Academy Press, Washington, DC (Panel on Confidentiality and Data Access, Committee on National Statistics).
- [8] Fienberg, S. E. (2000), ‘Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances’, paper presented at the IAOS

Conference on Statistics, Development and Human Rights, Montreux, Switzerland, 4 to 8 September 2000.

- [9] Fienberg, S., Makov, U. E. and Sanil, A. P. (1997), ‘A Bayesian approach to data disclosure: Optimal intruder behaviour for continuous data’, *Journal of Official Statistics*, Vol. 13, pp. 75–90.
- [10] Hurkens, C. A. J. and Tiourine, S. R. (1998), ‘Models and methods for the microdata protection problem’, *Journal of Official Statistics*, Vol. 14, pp. 437–447.
- [11] Lambert, D. (1993), ‘Measures of disclosure risk and harm’, *Journal of Official Statistics*, Vol. 9, pp. 313–333.
- [12] Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1999), ‘A method for data-oriented multivariate micro-aggregation’, *Statistical data protection, Proceedings of the Conference*, Lisbon, Portugal, Eurostat, pp. 88–89.
- [13] Subcommittee on Disclosure Avoidance Techniques (1978), ‘Report on statistical disclosure and disclosure-avoidance techniques’, Statistical Working Paper 2, Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, US Department of Commerce, Washington, DC.
- [14] Trottini, M. (2001), ‘Users’ uncertainty, disclosure risk and data utility’, Technical Report 2-2001, Departamento de Estadística y Investigación Operativa, Universitat de Valencia.
- [15] Willenborg, L. and de Waal, T. (2001), ‘Elements of statistical disclosure control’, *Lecture Notes in Statistics*, Vol. 155, Springer-Verlag, New York.

Uniqueness, urn models and disclosure risk

Stephen E. Fienberg (*) and Udi E. Makov (**)

(*) *Department of Statistics, Carnegie Mellon University, United States*

(**) *Department of Statistics, University of Haifa, Israel*

Keywords: Bayesian models, contingency tables, disclosure risk assessment, genetic evolution, log-linear models, Polya urn model, confidentiality

Abstract

The prevalence of categorical observations that are unique in a sample and also still unique in the population is usually taken as the measure of the overall risk of disclosure in the sample data. Samuels (1998) suggested adopting evolutionary processes and their associated urn models as a framework for estimating this prevalence. We re-examine his proposal and suggest several extensions that arise naturally in the Bayesian statistical framework. We provide a brief report on some empirical studies using data provided by the Israel Central Bureau of Statistics. We also link this approach to ones based on the structure of cross-classifications allowing for differential, per-unit forms of risk assessment.

1. Introduction

When a sample unique is also a population unique identity disclosure becomes much more likely and hence a source of profound worry for data-gathering agencies. Many authors have attempted to estimate the prevalence of sample uniques which are population uniques in cross classifications of categorical variables, i.e., multiway contingency tables. As we noted in Fienberg and Makov (1998), methods based on a frequency of frequencies approach dominated this literature (e.g. see Bethlehem et al., 1990; Chen and Keller-McNulty, 1998; Skinner and Holmes, 1998; Skinner et al., 1994). More recently, however, a number of authors suggested a more structured approach using log-linear and logistic models (e.g. Fienberg and Makov, 1998, and Skinner and Holmes, 1998) which attempt to capture the underlying probability structure of the contingency table.

In a somewhat different vein, Samuels (1998), starting from Chen and Keller-McNulty (1998), suggested adopting evolutionary genetics-inspired processes utilising urn models as a framework for this problem, although his approach still builds on the traditional frequency of frequencies structure which treats cells with the same count exchangeably. In this paper, we first build upon Samuels' approach and extend it using simulation methodologies associated with the Bayesian approach that arises naturally in these problems from taking mixtures of distributions. After outlining some numerical studies using data provided by the Israel Central Bureau of Statistics, we suggest one way to adapt the evolutionary urn model approach to draw strength from the contingency table structure using log-linear models and thereby accord different cells containing sample uniques different probabilities of being unique in the population.

2. Urn models for genetic evolution

Geneticists have long studied processes for explaining genetic diversity, and in the case of evolution in the absence of selection, suggested several models for explaining the growth of a population by means of reproduction and mutation. In particular, they have modelled the partition structure of the various allelic states of a gene by means of Polya type urn models, where an urn containing coloured balls is used to govern the growth in the number of balls and their colour (e.g. see Hoppe, 1987). The analogy with categorical data is as follows: coloured balls each representing an entry in a multi-way contingency table and the numbers represent the counts.

By focusing on balls whose colour is singly represented, depicting unique entries in the sample, we allow the sample to grow to the size of the population, observing those uniquely coloured balls in the sample which remained uniquely coloured in the population. We start by considering a process, denoted by $\{X_n\}$, which we generate through a sequence of draws from an urn containing black balls and other non-black balls of various colours. Hence all balls are equally likely to be drawn from the urn. For convenience, we label the colours by natural numbers, with black being number 1 and the rest of the colours labelled sequentially as the need arises.

The sampling regime of the urn model is as follows. At any stage, say n , we draw a ball from the urn. If it is black, we replace it along with an additional ball of a previously unobserved colour. If we draw a non-black ball, we return it to the urn along with another ball of the same colour. The process $\{X_n\}$ records the colours of the additional balls added to the urn. If the process starts with only θ black balls then $X_1=1$, $X_2=1$ or 2 , $X_3=1, 2$, or 3 , etc. The sequence of draws results in a sequence of random partitions denoted by $\{\Pi_n\}$, where, for a given value of n , the partition Π_n is a vector $\mathbf{a}=(a_1, a_2, \dots, a_n)$, such that a_1 is the number of colours that appeared once, a_2 is the number of colours that appeared twice, etc. In the genetic context for a given n , this corresponds to a partition of a fixed sample of n selectively-equivalent genes into a number of different gene types (alleles), or in the case of a contingency table, it corresponds to a partition of the cells according to the frequency of the cell entries, i.e., the *frequency of frequencies*. Clearly, a_1 is the number of balls each having a unique colour, and this corresponds to the count of the number of cells in the contingency table which contain an entry of 1. The quantity $\sum_i a_i = k$ provides the number of different colours present in the urn or nonzero cells in the contingency table, and $\sum_i ia_i = n$ is the sample size. In this partition we ignore the black balls since we regard them as the means for generating all other coloured balls or introducing mutations (new coloured balls).

As an illustration, we take an example analysed by Chen and Keller-McNulty (1998) (see also Samuels, 1998) involving a sample of 87 959 elements cross-classified according to five categorical variables, taken from the 1980 decennial census. In this sample $a_1=222$ coloured balls appeared once, corresponding to 222 unique cells, $a_2=111$ coloured balls appeared twice, corresponding to cells with two entries, etc. Finally, $a_{3649}=1$ corresponds to

a single cell with 3 649 entries. For this example, there are $k = \sum_i a_i = 1\,024$ non-empty cells, and a sample size of $n = \sum_i i a_i = 87\,959$.

Given a sample size n , the distribution of any partition, \mathbf{a} , is given by

$$P[\Pi_n = \mathbf{a} | n, \theta] = \frac{n!}{[\theta]^n} \prod_{i=1}^n \frac{\theta^{a_i}}{i^{a_i} a_i!}, \quad (1)$$

where $[\theta]^n = \theta(\theta+1) \dots (\theta+n-1)$. As a consequence of (1) we can easily establish that the probability distribution of the number of distinct colours k given n is

$$P(k | n, \theta) = |S_n^k| \theta^k / [\theta]^n, \quad (2)$$

where S_n^k is a Stirling number of the first kind. The conditional distribution of \mathbf{a} given k and n , is then

$$P(\mathbf{a} | k, n) = \frac{n!}{|S_n^k|} \prod_{i=1}^n i^{a_i} a_i!. \quad (3)$$

Equation (1) is known as Ewens' sampling formula in population genetics (conjectured by Ewens, 1972, and established by Karlin and McGregor, 1972). The derivation of this formula was inspired by the non-Darwinian theory of evolution. Under this theory, the genetical variation is not due to natural selection but arises as a result of purely stochastic changes in gene frequencies. Equation (1) provides the partition distribution of a sample of n genes into various types when no selective differences are assumed between these types of genes. Kingman (1980) showed that this formula arises in various models, in all of which a population is genetically evolving through reproduction and mutation and it represents the limit distribution of the genetic content in a random sample taken from the population.

Ewens' formula is related to the GEM distribution, which is defined as follows. Suppose species (colours) have random frequencies $P = (P_1, P_2, \dots)$ satisfying $0 < P_i < 1$, $\sum_{i=1}^{\infty} P_i = 1$.

Now suppose further that $P_1 = w_1$, $P_r = (1 - w_1)(1 - w_2) \dots (1 - w_{r-1})w_r$, where the w_1, w_2, \dots are i.i.d., according to a $Beta(1, \theta)$ density. The decreasing order statistics $(P_{(1)}, P_{(2)}, \dots)$ have the *Poisson-Dirichlet* distribution with parameter θ (Kingman, 1975). In a sample of n species (colours) from this population, the partition of the various species (colours) is distributed according to Ewens' formula.

If we assume that the partition density in equation (1) is correct, then the expected proportion of sample uniques which are population uniques is

$$\frac{n + \theta - 1}{N + \theta - 1}, \quad (4)$$

(see Samuels, 1998, for details). Thus for a given sample size, for inferences focus on the estimation of θ . From Ewens (1972), we know that the method of maximum likelihood produces an approximate estimator, $\hat{\theta}_{MLE}$, which is the solution of

$$k = \theta \ln \left[1 + \frac{n}{\theta} \right]. \quad (5)$$

Samuels (1998) applied this model to the Chen and Keller-McNulty (1998) data and he found it inadequate, with higher proportions of population uniques than estimation with equation (5) would suggest. He modified the model by introducing an additional parameter M as follows. The sampling process starts with θ black balls and additional M balls of 'primary' colours, whose role is similar to those of the black balls, i.e., they are not counted as part of the sample or the population. If any of the M balls are drawn, however, they are replaced and another ball of the same colour is added. Essentially, the M coloured balls, which we do not count in the partition, help give the coloured balls a head start and thus should improve the estimate of the proportion of sample uniques that are unique in the population. The expected proportion of sample uniques which are population uniques is

$$\frac{n + M + \theta - 1}{N + M + \theta - 1}. \quad (6)$$

In his empirical examples, Samuels (1998) shows that this extended model does improve the resulting estimate of the fraction of sample uniques that are also population uniques for small sampling fractions, e.g. $f=n/N$ on the order of 0.1. But as f grows, the model begins to increasingly underestimate the proportion. Thus we have empirical evidence for at least some actual official statistics data releases, that we need to incorporate the sampling fraction into our model somehow, perhaps through θ directly. We return to the issue of utilising the sampling fraction below.

3. Bayesian extensions to urn model structure

Because of their simple structure, we can't expect the urn models to represent the actual process governing the recreation of a population. We can modify them, however, and we can make them sufficiently flexible to produce potentially useful results.

The posterior distribution of θ , given the sufficient statistic k , is given by

$$h(\theta|k, n) \propto \frac{\theta^k}{[\theta]^n} g(\theta), \quad (7)$$

where $g(\theta)$ is a prior distribution of θ . A simple conjugate prior distribution for this likelihood does not seem to exist; however, since the Ewens' distribution of equation (1) belongs to the exponential family, it is likely that it will lend itself to adaptive rejection sampling of the kind reported in Smith and Gelfand (1992) and Gilks (1992), and hence we can investigate the posterior distribution in (7) by means of simulation. We can achieve additional flexibility if we choose the prior distribution to be a mixture of the type $\pi g_1(\theta) + (1-\pi)g_2(\theta)$. In this case the posterior expectation of the proportion of uniques in the sample which are also population uniques is given, approximately, by

$$\frac{n + \pi E_1(\theta) + (1 - \pi) E_2(\theta) - 1}{N + \pi E_1(\theta) + (1 - \pi) E_2(\theta) - 1}, \quad (8)$$

where E_i is the expectation with respect to g_i . By letting $M + \theta = \pi E_1(\theta) + (1 - \pi) E_2(\theta)$ in (6), we obtain (8) and hence the prior mixture provides the additional flexibility of the two-parameter model discussed in Samuels (1998).

Samuels takes a single sample of size n and uses it for estimating θ . This could be sufficient if the urn model were accurate. Since this is not the case and since the methodology is more reliable the larger is n (this is common to other methodologies in the uniqueness context) it may be advisable to collect data from several samples and incorporate the combined information for estimating θ . This can be done by artificially creating such samples, i.e. by taking samples of varying size, without replacement, from a population or from a particularly large sample. Suppose we have at hand q such samples, $(\mathbf{a}^1, n_1), \dots, (\mathbf{a}^q, n_q)$, where the pair (\mathbf{a}^i, n_i) represents the partition of the i th sample of size n_i .

Analysis now relies on a richer likelihood incorporating all q of the samples:

$$\ell(\theta | n_1, \dots, n_q) = \prod_{i=1}^q \frac{\theta^{k_i}}{[\theta]^{n_i}}, \quad (9)$$

where k_i is the number of distinct colours in the i th sample of size n_i . Bayesian analyses now work with the posterior distribution:

$$h(\theta | k_1, \dots, k_q, n_1, \dots, n_q) \propto \prod_{i=1}^q \frac{\theta^{k_i}}{[\theta]^{n_i}} g(\theta). \quad (10)$$

Using equation (10), we can obtain a value of θ which has an 'average' property since it attempts to account for values of θ suitable for samples constituting small and large sample fractions.

Based on Samuels' (1998) empirical work and our own, we conjecture that, for every sample fraction $f_n = n/N$, there should be a different value of θ for which this methodology produces reasonable results. We therefore propose to replace θ as a single parameter by a multi-parameter function, say $\gamma(\theta, \tau)$, where τ is a vector of training parameters, e.g.

$$\gamma(\theta, \tau) = \theta^{\lfloor 2 - f_n^\beta \rfloor^\delta}, \quad (11)$$

which converges to θ as $n \rightarrow N$, or the sampling fraction $f_n \rightarrow 1$, at a rate which is dictated by β and δ . We can exploit the new structure $\gamma(\theta, \tau)$ to estimate θ and τ either by using a subsampling approach or through a Bayesian approach.

3.1 Estimation through subsampling

In the subsampling approach we repeatedly subsample from the released sample of size n to produce m replicates or independent subsamples for each of q subsample sizes n_1, n_2, \dots, n_q , with corresponding sampling fractions $f_{n_1}, f_{n_2}, \dots, f_{n_q}$. Consider subsample size n_i . We compare each of the m replicate subsamples to the full released sample of size n from which it was drawn, and we record the fraction of unique individuals in the

subsample remaining unique in the full released sample, say $\hat{p}_{n_{ij}}$ for $j=1,2,\dots,m$. If Ewens' urn model were correct then we could compute the value of $\theta_{n_{ij}}$, say $\hat{\theta}_{n_{ij}}$ which would produce a fraction $\hat{p}_{n_{ij}}$ from equation (4), namely,

$$\hat{p}_{n_{ij}} = \frac{n + \hat{\theta}_{n_{ij}}}{N + \hat{\theta}_{n_{ij}} + 1}, \quad (12)$$

yielding

$$\hat{\theta}_{n_{ij}} = \frac{\hat{p}_{n_{ij}}(N-1) - (n-1)}{1 - \hat{p}_{n_{ij}}}, \quad (13)$$

for $i=1,2,\dots,q$, and $j=1,2,\dots,m$. According to Ewens' model, however, we can also compute the approximate maximum likelihood estimate of $\theta_{n_{ij}}$ finding the solution of equation (5) which we denote by $\hat{\theta}_{n_{ij}}^*$. We now choose $\hat{\tau}$, namely $\hat{\alpha}$ and $\hat{\delta}$, so that $\gamma(\hat{\theta}_{n_{ij}}^*, \hat{\tau})$ is closest in some sense to $\hat{\theta}_{n_{ij}}^*$. Least-squares is an obvious option. We can do this separately for each subsampling fraction, f_{n_i} , or in some joint fashion with respect to both replicates and the q different subsampling fractions.

Alternatively we can work directly in the probability scale, comparing the values of the empirically observed proportion $\hat{p}_{n_{ij}}$ with the estimated probabilities from equation (4) using the approximate maximum likelihood value for $\theta_{n_{ij}}$, $\hat{\theta}_{n_{ij}}^*$. Again, least-squares is an obvious option. We can do this separately for each subsampling fraction, f_{n_i} or in some joint fashion with respect to both replicates and the q different subsampling fractions.

The preceding subsampling method is based on 'backward evaluation' to study the mechanism which changes the partition of the data in the multi-way table as the sample fraction decreases. As an alternative we can follow a 'forward evaluation' approach, generating m sequences of subsamples of increasing size by literally drawing successive individuals from the released sample of size n without replacement. We will report details on this approach in a full version of this paper.

Our ultimate aim in both subsampling approaches remains the same: to estimate the curve relating the sample fractions with the actual estimates of probability that unique individuals in the subsample are also unique in the entire unreleased data held by the agency.

3.2. Estimation via Bayesian updating

The posterior distribution of θ now takes the form

$$h(\theta | k_1, \dots, k_q, n_1, \dots, n_q) \propto \int \left\{ \prod_{i=1}^q \frac{\gamma(\theta, \tau)^{k_i}}{[\gamma(\theta, \tau)^{k_i}]^{n_i}} g(\theta) \right\} f(\tau) d\tau, \quad (14)$$

where $f(\tau)$ is a prior on the parameters of the γ function. However, we need to obtain estimates of τ and this will be done through the joint posterior distribution

$$h(\theta, \tau | k_1, \dots, k_q, n_1, \dots, n_q) \propto \prod_{i=1}^q \frac{\gamma(\theta, \tau)^{k_i}}{[\gamma(\theta, \tau)^{k_i}]^{n_i}} g(\theta) f(\tau). \quad (15)$$

For estimation of θ and τ we use the Gibbs sampler (e.g. see Smith and Gelfand, 1992).

We can use $\hat{\theta}$ and $\hat{\tau}$ for estimating the expected proportion of sample uniques which are population uniques as follows:

$$\frac{n + \gamma(\hat{\theta}, \hat{\tau}) - 1}{N + \gamma(\hat{\theta}, \hat{\tau}) - 1}. \quad (16)$$

As an example, if we substitute equation (11) into equation (16), we obtain

$$\frac{n + \hat{\theta}^{\left[2 - f_n^{\hat{\beta}}\right]^{\hat{\delta}}} - 1}{N + \hat{\theta}^{\left[2 - f_n^{\hat{\beta}}\right]^{\hat{\delta}}} - 1}, \quad (17)$$

clearly demonstrating how the sample size influences the estimation process by means of $\hat{\beta}$ and $\hat{\delta}$, and the sampling fraction, f_n .

A more coherent Bayesian treatment avoids plug-in estimates and substitutes simulated values for β and δ generated from their respective marginal posterior distributions. Thus, we obtain the posterior distribution of the expected probability that a sample unique is also a population unique, and hence have a mechanism for providing additional information concerning the variability associated with estimating this probability.

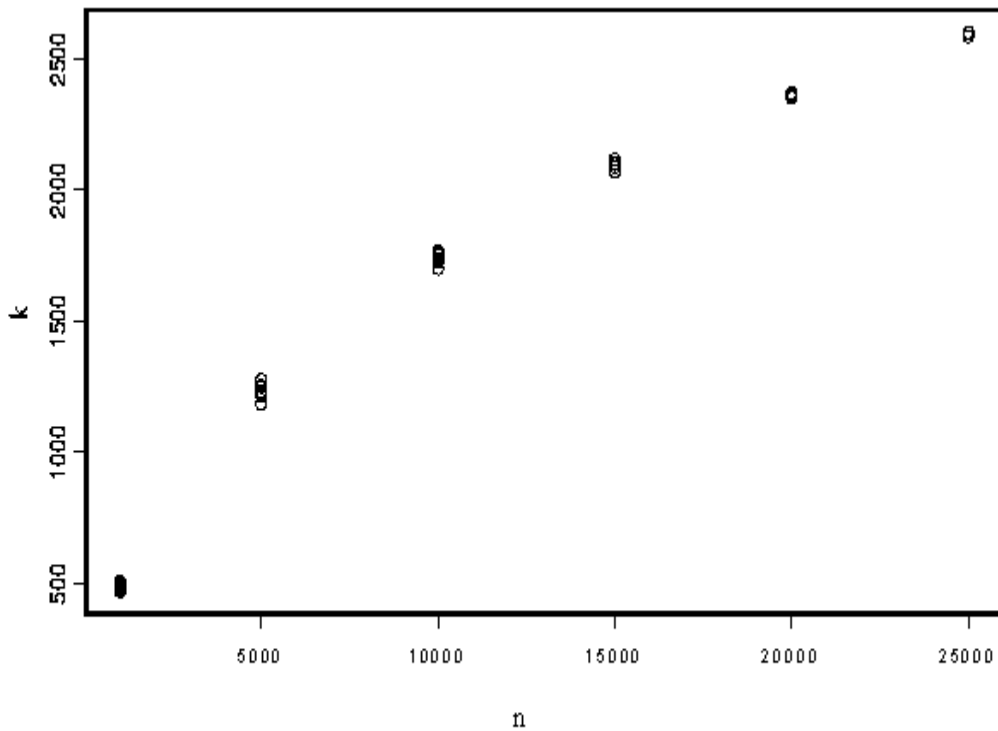
4. Numerical studies

Our numerical studies utilise 60 000 records taken from the 1996 labour force survey gathered by the Israeli Central Bureau of Statistics. For our experiments we assume that the released records consist of five chosen variables forming a multi-way table with 26 400 cells. We subsampled the 60 000 records forming six subsamples N_1, N_2, \dots, N_6 representing potential population sizes. From each of these 'populations', we took subsamples n_1, n_2, \dots, n_q representing the released samples. We replicated subsampling n out of N 3 times. The

following table provides the detailed design of the experiment.

Population size	Released sample size	Fraction	Fraction's symbol
N	n		
5 000	1 000	0.2	E
10 000	1 000	0.1	D
10 000	5 000	0.5	I
15 000	1 000	0.067	C
15 000	5 000	0.334	K
15 000	10 000	0.667	K
20 000	1 000	0.05	B
20 000	5 000	0.25	F
20 000	10 000	0.5	I
20 000	15 000	0.75	L
25 000	1 000	0.04	A
25 000	5 000	0.2	E
25 000	10 000	0.4	H
25 000	15 000	0.6	J
25 000	20 000	0.8	M
30 000	25 000	0.834	N

As the released sample size n increases one expects an increase in k , the number of 'colours', corresponding to the number of non-empty cells in the multi-way table. From a graph of k vs. n for all subsamples (Figure 1), the diminishing rate of the increase is obvious. The intrinsic curve relating k to n is a function of the data and the underlying model explaining it.

Fig 1: k vs. n

As k increases the probability that a sample unique is also a population unique increases. This is clearly shown in Figure 2 where P_{usup} represents the measured proportion of sample uniques which are also population uniques. ‘Sign=fraction’ indicates that letters appearing in the figure correspond to sample fractions.

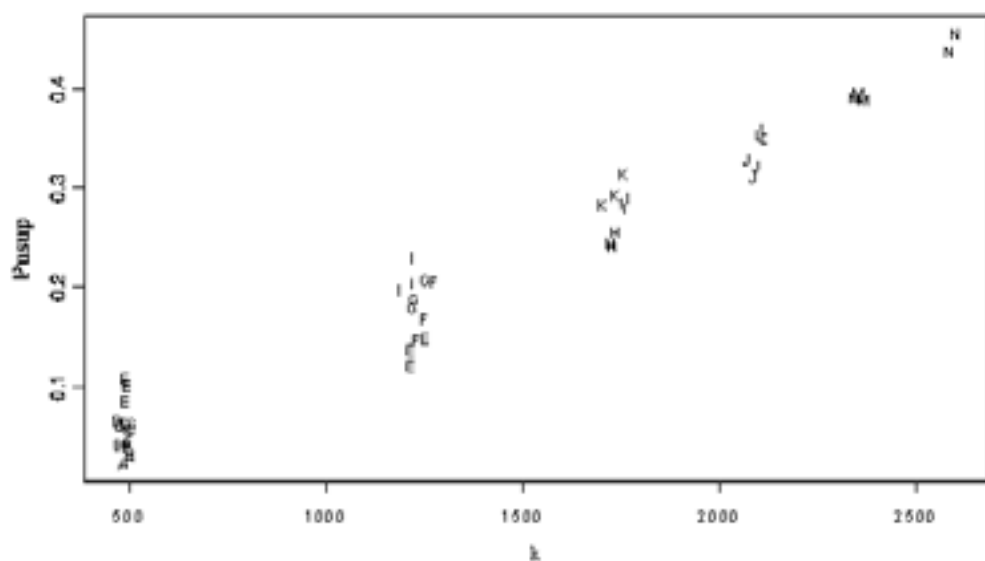


Fig 2: Pusup vs. k (Sign=fraction)

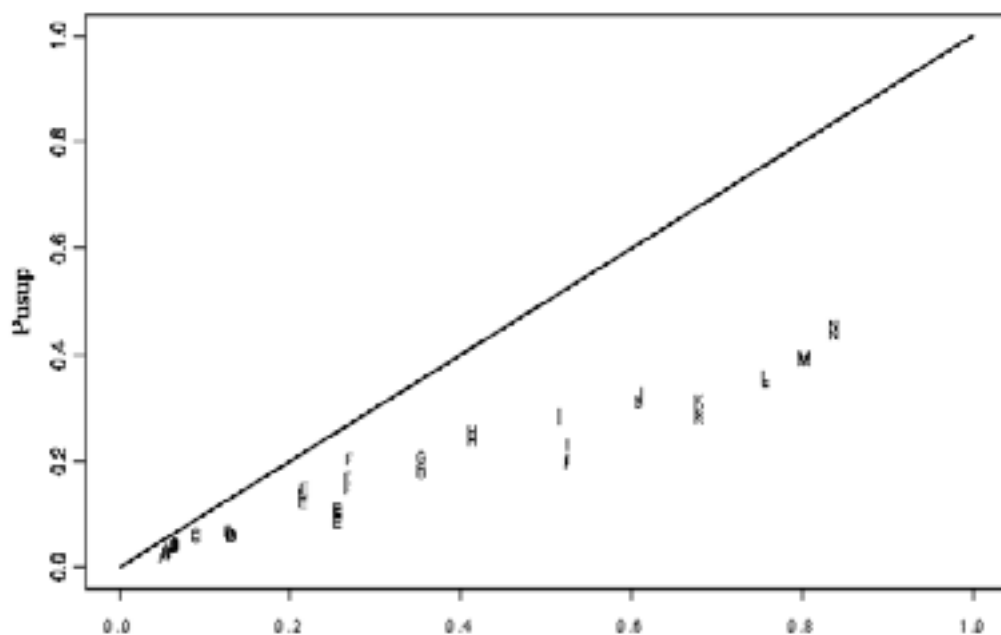


Fig 3: Pusup vs. p-hat (Sign=fraction)

If the urn model were correct the expected proportion of sample uniques which are population uniques, given in (4), was supposed to match, to some degree of accuracy, the

measured proportion of sample uniques which are also population uniques. Figure (3) clearly demonstrates the failure of the urn model. Here \hat{P} denotes the estimated expected proportion of sample uniques which are population uniques, obtained by substituting the MLE of θ into (4). The marked underestimation demonstrated in Fig. (3) clearly cannot be rectified by modifying the value of θ as suggested in (13). This is demonstrated in Figure 4 where the vertical axes corresponds to the value of θ that would produce, via (4) or (13), P_{usup} .

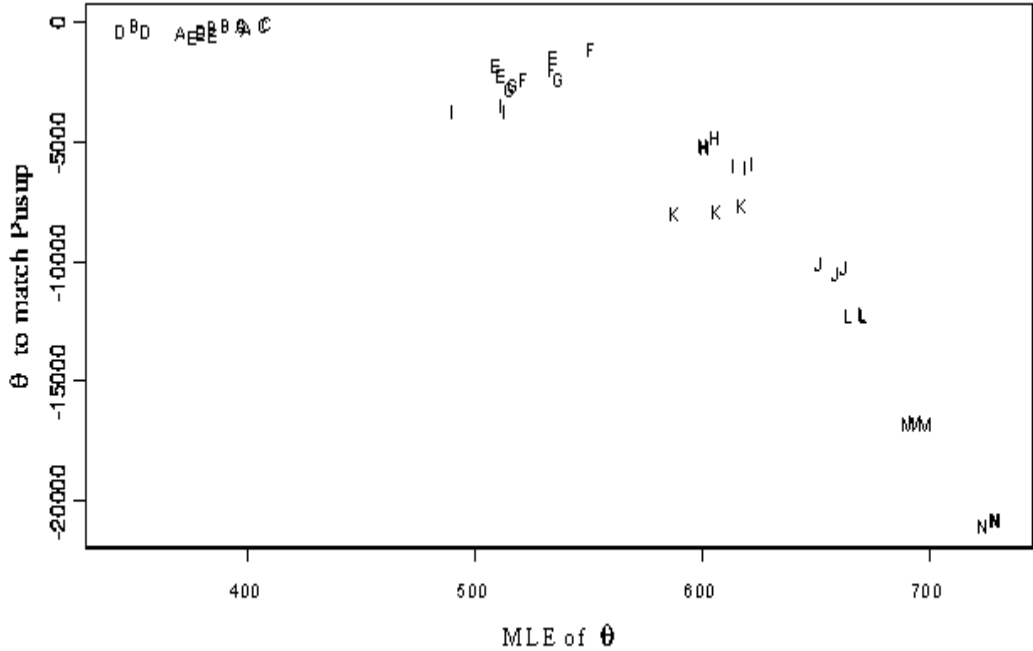


Fig 4: θ to match P_{usup} vs. MLE of θ (Sign=fraction)

Clearly, for most subsamples the use of (4) cannot be justified since it entails negative values of θ . Equation (4) can be modified to produce values close to those of P_{usup} as follows

$$\frac{n + \theta - 1}{N^* + \theta - 1}. \quad (18)$$

Here N^* is a modified population size to be estimated through the following equation

$$P_{usup} = \frac{n + \hat{\theta}_{MLE} - 1}{N^* + \hat{\theta}_{MLE} - 1}. \quad (19)$$

Solving for N^* ,

$$N^* = \frac{n + (1 - P_{u \sup}) (\hat{\theta}_{MLE} - 1)}{P_{u \sup}},$$

we plot, in figure 5, $NR = N^*/N$ vs. $\hat{\theta}_{MLE}$ for all subsamples.

Clearly, NR can be explained by the sample fraction, $\hat{\theta}_{MLE}$ and possibly the interaction between the two. Indeed, when a linear model was fitted to the data the $\hat{\theta}_{MLE}$, fraction and their interaction proved significant (p -values = 0.0000, 0.0000, 0.0029, respectively). The adequacy of the model is clearly demonstrated (we note the presence of suspected outliers) in its the residual plot (Figure 6).

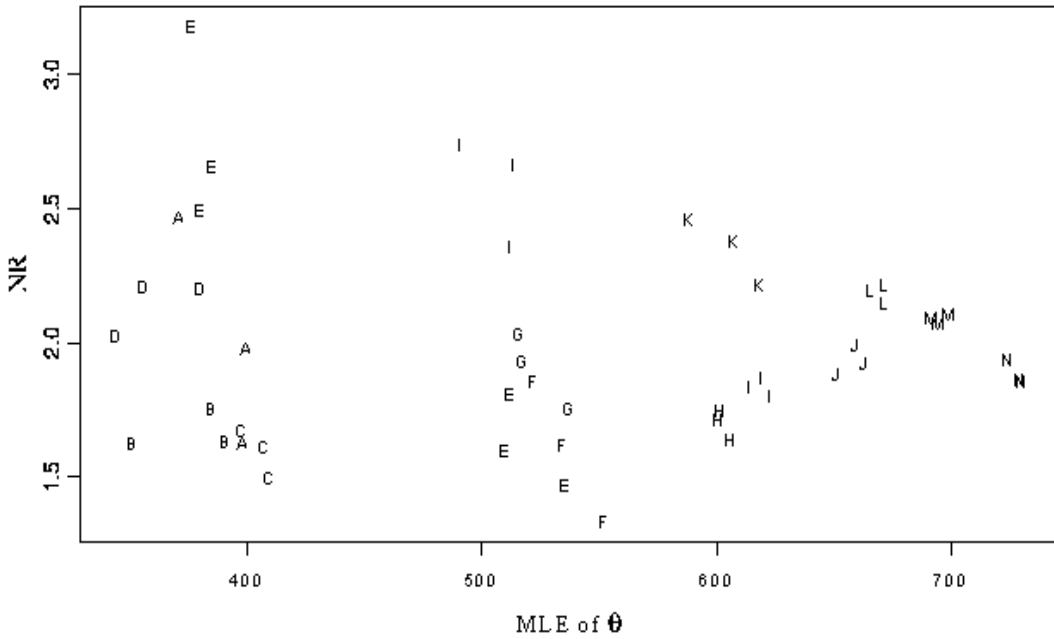


Fig 5: NR vs. MLE of θ (Sign=fraction)

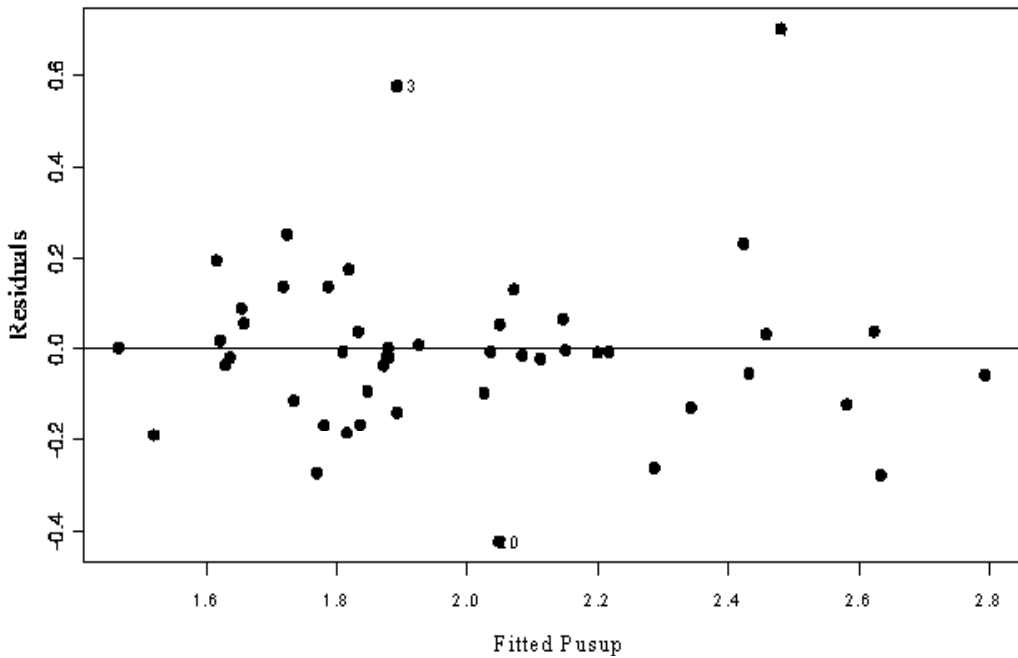


Fig 6: Residual Plot

The potential benefit of the model was examined as follows: A linear model was fitted using all subsamples such that $N \leq 15000$. Using the estimated parameters of the model and the actual values of the $\hat{\theta}_{MLE}$'s, the P_{usup} were predicted for the remaining subsamples ($N \geq 15\,000$). The result is shown in Figure 7 where the measured P_{usup} and the predicted P_{usup} are plotted. Clearly, though not sufficiently accurate, the results are promising.

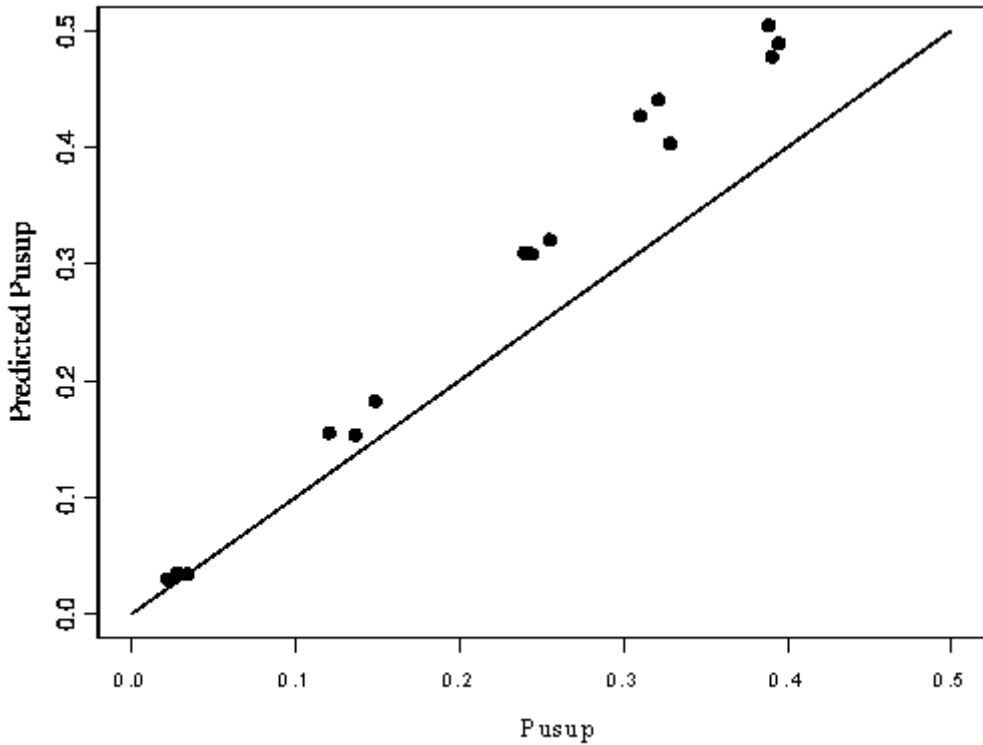
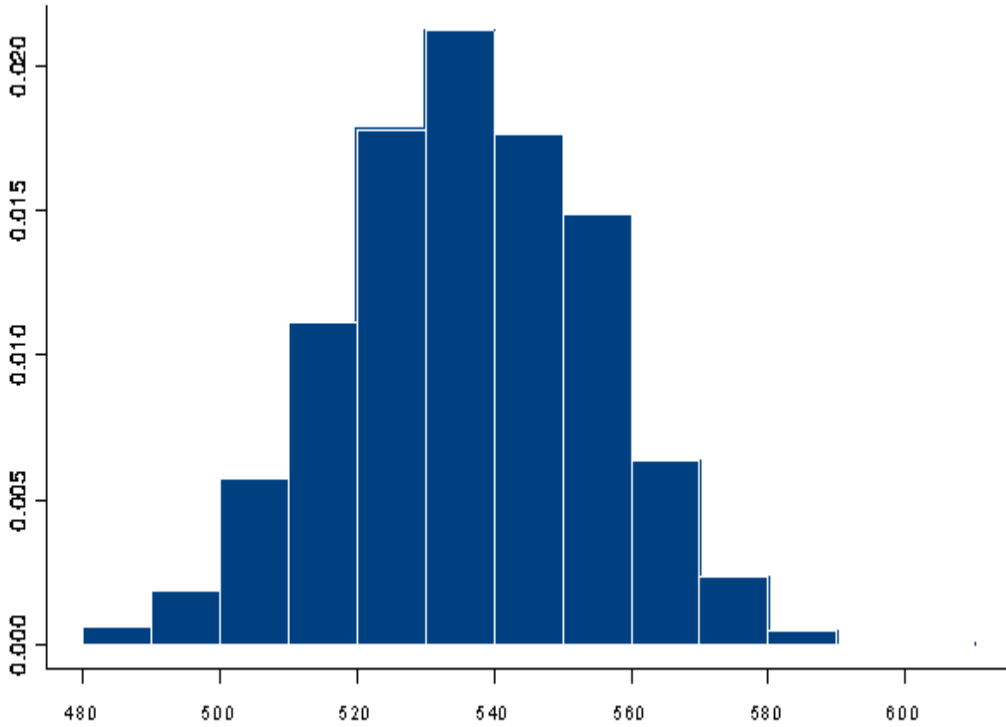


Fig 7: Predicted vs. measured Pusup

A full Bayesian analysis should include three phases:

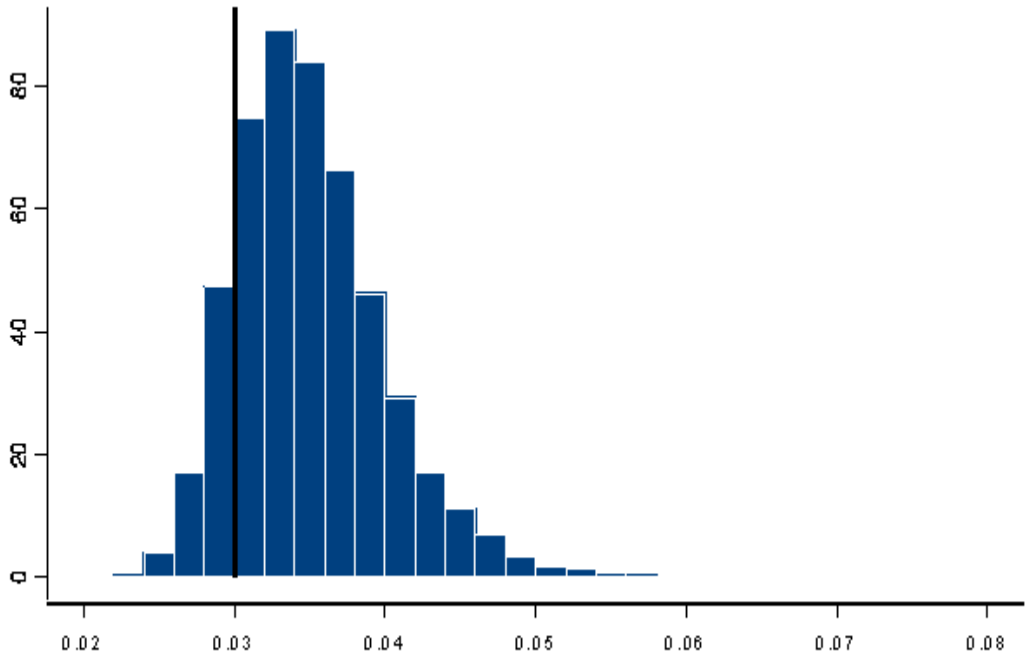
- The prediction of k via n . This would amount to estimating the curve shown in Figure 1.
- The prediction of θ based on the observed k 's.
- The prediction of P_{usup} via a linear model based on a given sample fraction and the predicted values of θ .

the second phase was attempted for $N=25\,000$ and $n=5\,000$. For a likelihood function based on (9), a normal prior distribution for θ was chosen with mean and variance empirically estimated using the values of the $\hat{\theta}_{MLE}$'s. Employing a rejection method, the posterior distribution of θ was obtained and shown in Figure 8.

Fig 8: Posterior distribution of θ

The third phase was attempted as follows:

The β s of the linear model were given a normal prior distribution and σ^2 , the variance of the noise, was given an inverse gamma prior distribution. Using the Gibbs sampler, posterior distributions of the β s were generated based on subsamples for which $N \leq 1500$ and then used to predict NR and P_{usup} via (19) for $N = 25000$ based on $n=1000$. Figure 9 shows the posterior distribution of the predicted P_{usup} .

Fig 9: Posterior distribution of predicted P_{usup}

The measured P_{usup} is indicated by a vertical line. Clearly, the distance between the actual P_{usup} and the mode of the posterior distribution is only about 0.05. Further research is needed to reduce this discrepancy.

5. Incorporating contingency table structure

To this point, the models we have been using treat cells with equal sample counts exchangeably, ignoring their actual placement in the contingency table in which they are situated. Thus Ewens' model treats all cells with zero counts alike in choosing to convert one of them to a sample unique whose cell count is 1, or all cells containing sample uniques alike in choosing one to convert to a cell with a count of 2. But, as the literature on log-linear models for contingency tables makes clear, all sample zeros are not all alike and they often correspond to cells with very different underlying probabilities, expressible in terms of marginal totals of different magnitudes (e.g. see Bishop et al., 1975). Thus we desire a model that combines some of the attractive features of the urn schemes for the frequencies of frequencies described earlier in this paper with something that can reflect the differential treatment of the log-linear models. In the previous sections, the urn models we considered treated all cells such that $f_i=j$ exchangeably as part of the partition function

a, and focused on the determination of the prevalence $\sum_{i=1}^K P(F_i = 1 / f_i = 1)$. Here we focus on the evaluation of $P(F_i | f_i)$ separately for each cell i , because we think in terms of a superpopulation for the cross-classification such that a log-linear model of the form

$$\log(\pi_i) = g(u_i) \quad (20)$$

describes the underlying structure of the population cell probabilities. Thus we can no longer treat the components of the prevalence as exchangeable. One way to think about doing this is to think in terms of urn models where the balls are assigned unequal probabilities of selection or weights as follows: assign each cell in the cross classification its own colour, and its own a priori probability. Assign each coloured ball a weight that is proportional to the posterior expected value under model (20), based on the balls that currently make up the urn. Begin the process with an urn containing θ black balls. Draw balls from the urn with unequal probabilities that are determined in proportion to their weights. When we draw a black ball, return it to the urn and add a new colour from a pool of colours of yet unrepresented cells drawn in proportion to the weights (estimated probabilities) associated with the yet unobserved cells. When we draw a coloured ball, return it to the urn with another ball of the same colour. Because this modified urn scheme actually builds on the log-linear model for the cell probabilities, it gains strength from the observed marginal totals corresponding to the minimal sufficient statistics, and it gives differential attention to the growth of cells as a function of the growth of the relevant marginals. Thus the cell-by-cell components of the prevalence are no longer treated exchangeably. This weighted urn model relies initially on the a priori assessment of cell probabilities and then uses Bayesian methods to update those probabilities as data accumulate. But it also utilises the mechanism of the black balls to insert an evolutionary structure to the composition of the urn. Further variations on this theme would allow not only for θ to vary as a function of the sampling fraction but also directly tie its magnitude to information about the underlying cell probabilities, $\{\pi_i\}$, for the cross-classification. Since all of these urn models rely on externally computed weights, we can easily substitute a more elaborate Bayesian model-averaging approach to allow for our lack of certainty about the appropriate log-linear model in equation (20). The challenge that faces us is the determination of the properties of such weighted urn models. We can investigate these properties directly through simulation.

6. Acknowledgements

This research was supported in part by the U.S. Bureau of the Census through a contract with Westat and Carnegie Mellon University, by NSF Grant EIA-9876619 to the National Institute of Statistical Sciences, and by the Israel Central Bureau of Statistics through a contract with the University of Haifa. We are grateful to A. Hazan for assistance.

7. References

- [1] Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990), ‘Disclosure control of microdata’, *Journal of the American Statistical Association*, Vol. 85, pp. 38–45.
- [2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P.W. (1975), *Discrete multivariate analysis: Theory and practice*, MIT Press, Cambridge, MA.
- [3] Chen, G. and Keller-McNulty, S. (1998), ‘Estimation of identification disclosure risk in microdata’, *Journal of Official Statistics*, Vol. 14, pp. 79–95.
- [4] Ewens, W. J. (1972), ‘The sampling theory of selectively neutral alleles’. *Theoretical Population Biology*, Vol. 3, pp. 87–112.
- [5] Fienberg, S. E. and Makov, U. E. (1998), ‘Confidentiality, uniqueness, and disclosure avoidance for categorical data’, *Journal of Official Statistics*, Vol. 14, pp. 385–397.
- [6] Gilks, W. R. and Wild, P. (1992), ‘Adaptive rejection sampling for Gibbs sampling’, *Applied Statistics*, Vol. 41, pp. 337–348.
- [7] Hoppe, F. M. (1987), ‘The sampling theory of neutral alleles and an urn model in population genetics’, *Journal of Mathematical Biology*, Vol. 25, pp. 123–159.
- [8] Karlin, S. and McGregor, J. L. (1972), ‘Addendum to a paper of W. Ewens’, *Theoretical Population Biology*, Vol. 3, pp. 113–116.
- [9] Kingman, J. F. C. (1975), ‘Random discrete distributions’, *Journal of the Royal Statistical Society, Series B*, Vol. 37, pp. 1–22.
- [10] Kingman, J. F. C. (1980), ‘Mathematics of Genetics Diversity’, CBMS-NSF Regional Conference Series in Applied Mathematics, 34, Society for Industrial and Applied Mathematics, Philadelphia.
- [11] Samuels, S. M. (1998), ‘A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment’, *Journal of Official Statistics*, Vol. 14, pp. 373–383.
- [12] Skinner, C. J. and Holmes, D. J. (1998), ‘Estimating the re-identification risk per record in microdata’, *Journal of Official Statistics*, Vol. 14, pp. 361–372.
- [13] Skinner, C. J., Marsh, C., Openshaw, S. and Wymer, C. (1994), ‘Disclosure control for census microdata’, *Journal of Official Statistics*, Vol. 10, pp. 31–51.
- [14] Smith, A. F. M. and Gelfand, A. E. (1992), ‘Bayesian statistics without tears: A sampling-resampling perspective’, *American Statistician*, Vol. 46, pp. 84–88.

Structural analysis of the abortion time series reported in the ISTAT summaries

Riccardo Bellazzi and Paolo Magni

Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Italy

Keywords: Bayesian analysis, structural time series analysis, dynamical systems, abortion, Markov-Chain Monte Carlo

Abstract

After the approval of the 1978 law on voluntary abortion in Italy, the Italian healthcare system allowed the practice of voluntary abortion before the third month of pregnancy. Since 1980, the Italian Institute of Statistics (ISTAT) has collected data on the abortion frequency per month and per administrative local area. Although a preliminary analysis of the data showed that, after an initial increase in the number of abortions, the number progressively decreased over the years, there is no insight into the existence of periodicity in the time series and into the local effects related to the regional habits and social environments. The aim of our study is therefore to extract local trends and periodicity from the data collected by ISTAT, by combining a 'structural model' of the time series and Bayesian statistics. This paper describes the stochastic model and its Bayesian estimation through a Markov-Chain Monte Carlo approach on the Italian abortion data.

In our analysis, the trend component is very regular and shows clearly how after an initial period of increase, after June–July 1983 the voluntary abortion trend decreases constantly until the end of the study. The periodic component shows an astonishing regularity too, suggesting that the Italian people have a seasonal preference for voluntary abortion.

1. Introduction

After the approval of the 1978 law on voluntary abortion in Italy, the Italian healthcare system allowed the practice of voluntary abortion before the third month of pregnancy. Since 1980, the Italian Institute of Statistics (ISTAT) has collected data on the abortion frequency per month and per administrative local area (province). It is interesting to analyse the accumulated data in order to provide some elements to the public discussion on a theme with significant ethnical implications. In particular, we extracted from the raw data: (i) the average course in order to extract change points; and (ii) the periodic components in order to highlight possible seasonal preferences for voluntary abortions. Finally, we also looked for regional similarities or differences in this practice in order to start a sociological analysis.

Several statistical techniques are available to study seasonal effects, but, in general, they are only able to extract a global trend and a global periodic component: that is, the best trend and the best periodic component over the whole period. Using moving windows or other techniques, such as Kalman filtering, it is possible to extract 'local' components. However, such approaches are often not flexible enough to follow the temporal pattern

of real data due to the fixed window length. If the time window is unnecessarily long, the trend may not capture complex patterns made of slow variations; on the other hand, if the time window is too short, the algorithm does not provide a sufficiently smooth trend estimate. Therefore, the choice of the time window turns out to be ad hoc.

In this paper, both the trend and periodic components are modelled as randomly varying over time. If the statistical properties of such components were known in advance, it should be possible to perform the structural analysis by means of the Kalman filter. In order to enhance the generality and usability of the method, we prefer to describe the rates of variation as random variables whose variances are treated as ‘hyperparameters’ within a Bayesian estimation framework. The computation of the posterior densities of the rates given the data involves intractable integrals that hamper an analytic approach. Therefore, we resort to a Markov chain Monte Carlo method that produces the desired estimate as the result of a stochastic simulation procedure.

2. The database

The database provided by ISTAT consists of two relational tables: the first contains the number of abortions per year, month and local area, and the second the encoding of the local areas. The first table has 16 880 records, corresponding to the years 1980–94 and to 95 local areas. We limited the analysis to the years 1980–94 because after that time several new local areas were added and inclusion/exclusion of these areas could bias the study. The database had 220 missing values, corresponding to 1.3 % of the data records.

In this paper, we will consider only the voluntary abortion time series (VATS) at a national level obtained by adding all the local data for each month. In this way, a complete time series of 180 sets of data was obtained. However, some data on VATS at the national level may underestimate the real number of the voluntary abortions performed because some local data are not available for some months. In general, this problem does not seem to be very important since the number of local areas is very high, while the missing data are distributed along the whole period under investigation.

3. The methodology

To analyse the VATS, we adapted a methodology that has been applied to the blood glucose time series in the first-type of diabetes monitoring (Bellazzi et al., 2000).

We assume that each of the data of the time series can be expressed as a sum of separate components, that represent its underlying structure. In the case of the VATS, three components are considered — a trend component t_i , a cyclic (or periodic) component c_i (with a period of one year) and a noise component v_i (taking into account the uncertainty of the data) — so that

$$y_i = t_i + c_i + v_i, \tag{1}$$

where v_i is the number of abortions (at the national level) at time i . However, because a generic periodic signal can be written through a Fourier series expansion, the periodic component c can be approximated with a finite sum of sine or cosine waves at appropriate frequency.

For reasons that will be clear later, it is convenient to rewrite equation (1) as a state–space discrete-time model, where the sampling times correspond to the monthly grid:

$$t_{i+1} = t_i + s_i \quad (2)$$

$$s_{i+1} = s_i \quad (3)$$

$$c_{i+1}^{(j)} = c_i^{(j)} \cos(2\pi j f) + r_i^{(j)} \sin(2\pi j f) \quad (4)$$

$$r_{i+1}^{(j)} = -c_i^{(j)} \sin(2\pi j f) + r_i^{(j)} \cos(2\pi j f), \quad j = 1, \dots, 4 \quad (5)$$

$$c_i = c_i^{(1)} + c_i^{(2)} + c_i^{(3)} + c_i^{(4)} \quad (6)$$

$$y_i = t_i + c_i + v_i, \quad i = 0, \dots, n-1 \quad (7)$$

with the initial conditions $t_0, s_0, c_0^{(1)}, r_0^{(1)}, c_0^{(2)}, r_0^{(2)}, c_0^{(3)}, r_0^{(3)}, c_0^{(4)}, r_0^{(4)}$, where s and r are auxiliary variables. In particular, s_i is the increment of the trend component between i and $i + 1$. In fact,

$$t_i = t_0 + i s_0.$$

Hence, the trend component is just a straight line, whereas the cyclic component is a linear combination of sine and cosine waves (equation (6)). In fact, the periodic component c is modelled through its truncated Fourier series, containing the fundamental harmonics and the higher-order harmonics with frequency $\leq 1/3$, in order to have at least three samples in one period. Since we look for cyclic components with a period equal to one year, while the data are available monthly, the fundamental frequency f should be taken as $1/12$ and the first four harmonics have to be considered.

The model (equations (2–7)), which is deterministic in the state equations, is not able to capture the variability of the time series, which can present trend and cyclic changes within a few years (or a few months). A suitable approach, that allows for both trend and cyclic variations, involves a stochastic description of the abortion dynamics. This is obtained by adding random noise to both the trend and cyclic components in equations (2–5). In particular, we used the following model:

$$t_{i+1} = t_i + s_i \quad (8)$$

$$s_{i+1} = s_i + w_{1i} \quad (9)$$

$$c_{i+1}^{(j)} = c_i^{(j)} \cos(2\pi j f) + r_i^{(j)} \sin(2\pi j f) \quad (10)$$

$$r_{i+1}^{(j)} = -c_i^{(j)} \sin(2\pi j f) + r_i^{(j)} \cos(2\pi j f), \quad j = 1, \dots, 3 \quad (11)$$

$$c_{i+1}^{(4)} = c_i^{(4)} \cos(2\pi 4 f) + r_i^{(4)} \sin(2\pi 4 f) \quad (12)$$

$$r_{i+1}^{(4)} = -c_i^{(4)} \sin(2\pi 4 f) + r_i^{(4)} \cos(2\pi 4 f) + w_{2i} \quad (13)$$

$$c_i = c_i^{(1)} + c_i^{(2)} + c_i^{(3)} + c_i^{(4)} \quad (14)$$

$$y_i = t_i + c_i + v_i, \quad i = 0, \dots, n-1 \quad (15)$$

with the initial conditions $t_0, s_0, c_0^{(1)}, r_0^{(1)}, c_0^{(2)}, r_0^{(2)}, c_0^{(3)}, r_0^{(3)}, c_0^{(4)}, r_0^{(4)}$, where w_{1i}, w_{2i}, v_i and the initial conditions are (a priori) independent random variables. w_{1i} and w_{2i} describe the random fluctuations of the trend and the periodic components. For simplicity, random fluctuations of the periodic component in this model are allowed only in the highest of the considered harmonics. In principle, we could add random variables to every harmonic, but the complexity of the stochastic model is likely to increase without providing appreciable benefits.

Following standard notation, we can rewrite the dynamic model (equations (8–15)) as

$$x_{i+1} = Fx_i + Gw_i \quad (16)$$

$$y_i = Hx_i + v_i \quad (17)$$

where

$$x_i = [t_i \ s_i \ c_i^{(1)} \ r_i^{(1)} \ c_i^{(2)} \ r_i^{(2)} \ c_i^{(3)} \ r_i^{(3)} \ c_i^{(4)} \ r_i^{(4)}]^T,$$

$$w_i = [w_{1i} \ w_{2i}]^T,$$

$$H = [1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0],$$

$$F = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\left(\frac{2\pi}{12}\right) & \sin\left(\frac{2\pi}{12}\right) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\sin\left(\frac{2\pi}{12}\right) & \cos\left(\frac{2\pi}{12}\right) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos\left(\frac{2\pi}{6}\right) & \sin\left(\frac{2\pi}{6}\right) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\sin\left(\frac{2\pi}{6}\right) & \cos\left(\frac{2\pi}{6}\right) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cos\left(\frac{2\pi}{4}\right) & \sin\left(\frac{2\pi}{4}\right) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\sin\left(\frac{2\pi}{4}\right) & \cos\left(\frac{2\pi}{4}\right) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cos\left(\frac{2\pi}{3}\right) & \sin\left(\frac{2\pi}{3}\right) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\sin\left(\frac{2\pi}{3}\right) & \cos\left(\frac{2\pi}{3}\right) \end{bmatrix}^T,$$

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T,$$

with the initial condition $x_0 = [t_0 \ s_0 \ c_0^{(1)} \ r_0^{(1)} \ c_0^{(2)} \ r_0^{(2)} \ c_0^{(3)} \ r_0^{(3)} \ c_0^{(4)} \ r_0^{(4)}]^T$.

In order to complete the stochastic model specification, it is necessary to assign the distribution of all the random variables. Herein, we assume that

$$\begin{aligned} p(w_{1i}) &= N(0, \sigma_{w_1}^2), \\ p(w_{2i}) &= N(0, \sigma_{w_2}^2) \\ p(x_0) &= N(0, \text{diag}([\sigma_{x_{01}}^2 \ \sigma_{x_{02}}^2 \ \sigma_{x_{03}}^2 \ \sigma_{x_{04}}^2 \ \sigma_{x_{05}}^2 \ \sigma_{x_{06}}^2 \ \sigma_{x_{07}}^2 \ \sigma_{x_{08}}^2 \ \sigma_{x_{09}}^2 \ \sigma_{x_{010}}^2])) \\ p(v_i) &= N(0, \sigma_{v_i}^2) \end{aligned}$$

where $N(\cdot, \cdot)$ denotes the normal distribution and diag is the diagonal matrix.

Unlike the approach for standard Kalman filtering, we consider variances $\sigma_{w_1}^2$ and $\sigma_{w_2}^2$ as unknown variables too; they are ‘hyperparameters’ to be estimated from data. We assume that the unknown variances have an inverse gamma distribution with known parameters (see Magni et al., 1998, for a detailed discussion). In this analysis, the parameters of the inverse gamma distribution are chosen in such a way that the prior distribution of the hyperparameters is sufficiently flat: in this sense, the final estimates are relatively insensitive to the tuning of the inverse gamma distributions. On the other hand, we assume that the statistics of the measurement error v_i are known completely, i. e. we consider the $\sigma_{v_i}^2$ as fixed parameters. Our approach can easily be generalised to cope with the problem in which the statistic model of measurement error is not completely known (Magni et al., 1998).

Starting from equations (16, 17), we can rewrite the dynamic model in the following static model, in order to estimate trend and periodic components easily:

$$y = LAz + v \quad (18)$$

$$x = Bz \quad (19)$$

where

$$z = [t_0 \ s_0 \ c_0^{(1)} \ r_0^{(1)} \ c_0^{(2)} \ r_0^{(2)} \ c_0^{(3)} \ r_0^{(3)} \ c_0^{(4)} \ r_0^{(4)} \ w_{1_0} \ w_{2_0} \ \dots \ w_{2_{n-2}} \ w_{2_{n-2}}]^T,$$

$$x = [t_0 \ c_0 \ \dots \ t_{n-1} \ c_{n-1}]^T,$$

$$A = \begin{bmatrix} H & 0 & \dots & 0 \\ HF & HG & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ HF^{n-1} & HF^{n-2}G & \dots & HG \end{bmatrix},$$

$$B = \begin{bmatrix} M & 0 & \cdots & 0 \\ MF & MG & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ MF^{n-1} & MF^{n-2}G & \cdots & MG \end{bmatrix},$$

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

and $v = [v_0 \dots v_{n-1}]^T$. The matrix L is introduced to manage missing data (Magni et al., 1998); when there are no missing data, L is an $n \times n$ identity matrix.

Given the stochastic model (equations (18, 19)), the Bayesian point estimation problem involves the computation of the first and second moments of the joint posterior probability distribution:

$$(20) \quad p(t, c, \sigma_{w1}^2, \sigma_{w2}^2 \mid y)$$

$$\text{where } t = [t_0 \dots t_{n-1}]^T, c = [c_0 \dots c_{n-1}]^T, y = [y_0 \dots y_{n-1}]^T.$$

This problem cannot be solved in a closed form, and thus we use Markov chain Monte Carlo (MCMC) methods. MCMC methods are based on two steps: a Markov chain and a Monte Carlo integration. By sampling from suitable probability distributions, we generate a Markov chain that converges (in distribution) to the target distribution, i.e. the distribution to be integrated. Then, we calculate the expectation through the Monte Carlo integration over the obtained samples. In this paper, we use the well-known Gibbs sampling scheme, proposed originally in Geman and Geman (1984).

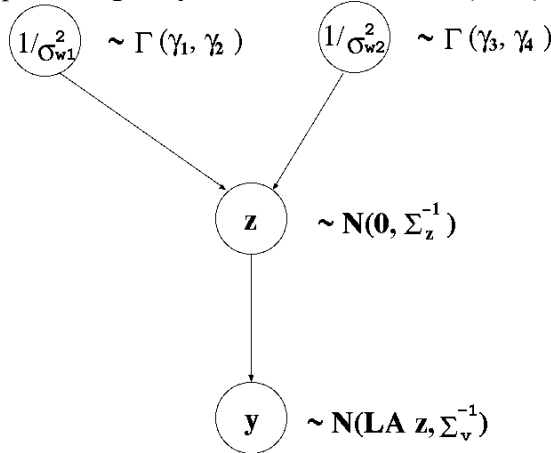


Figure 1: Bayesian model of the time series

Given the stochastic model (equation (18)), schematised in Figure 1, it is possible to derive the full conditional distributions necessary to run the Gibbs sampler estimator:

$$p\left(\frac{1}{\sigma_{w_1}^2} \mid \sigma_{w_2}^2, z, y\right) = \Gamma\left(n/2 + \gamma_1, (w_1^T w_1)/2 + \gamma_2\right) \quad (21)$$

$$p\left(\frac{1}{\sigma_{w_2}^2} \mid \sigma_{w_1}^2, z, y\right) = \Gamma\left(n/2 + \gamma_3, (w_2^T w_2)/2 + \gamma_4\right) \quad (22)$$

$$\begin{aligned} p(z \mid \sigma_{w_1}^2, \sigma_{w_2}^2, y) &= N\left(D^{-1}(LA)^T \Sigma_v^{-1} y, D^{-1}\right) \\ D &= (LA)^T \Sigma_v^{-1} (LA) + \Sigma_z^{-1} \end{aligned} \quad (23)$$

where $\Gamma(\cdot, \cdot)$ is the gamma distribution, $w_1 = [w_{10} \dots w_{1n-2}]^T$, $w_2 = [w_{20} \dots w_{2n-2}]^T$, $\Sigma_v = \text{diag}([\sigma_{v_0}^2 \dots \sigma_{v_{n-1}}^2])$, $\Sigma_z = \text{diag}([\sigma_{x_{01}}^2 \sigma_{x_{02}}^2 \dots \sigma_{x_{010}}^2 \sigma_w^2])$ and σ_w^2 is a vector containing the sequence $\{\sigma_{w_1}^2, \sigma_{w_2}^2\}$ repeated $n-1$ times. $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are the parameters of the prior gamma distributions of the variables $1/\sigma_{w_1}^2$ and $1/\sigma_{w_2}^2$.

From the samples of $p(\sigma_{w_1}^2, \sigma_{w_2}^2, z \mid y)$ drawn iteratively by the Gibbs Sampler using equations (21–23), it is straightforward to derive through equation (19) the sampling posterior distribution of the trend and cyclic components ($p(t, c, \sigma_{w_1}^2, \sigma_{w_2}^2 \mid y)$).

The adopted scheme provides for drawing samples of z (the vector containing the initial state value and all the other stochastic components of the system, i.e. w_1, w_2) from a multinormal distribution of $2(n-1) + 10$ dimensions. This implies that when n is very large it is computationally expensive to invert the matrix D , and to extract samples from a high-dimensional multinormal distribution. To solve these problems, it is possible to partition the set of the stochastic parameters in a different way or to resort to more general simulation strategies involving dynamic linear models (Carlin et al., 1992; Carter and Kohn, 1994; West and Harrison, 1997).

For the analysis of VATS, we took $\gamma_1 = 1, \gamma_2 = 10, \gamma_3 = 1, \gamma_4 = 0.1, \sigma_{x_{0i}}^2 = 10^8$ and $\sigma_{v_i}^2 = 30000$. By choosing large values for $\sigma_{x_{0i}}^2$, the posterior distribution of the initial conditions will be driven only by the data.

4. Results

The official data collected from ISTAT about the voluntary abortions in Italy on a monthly basis for the period from 1980 to 1994 are reported in Figure 2(a).

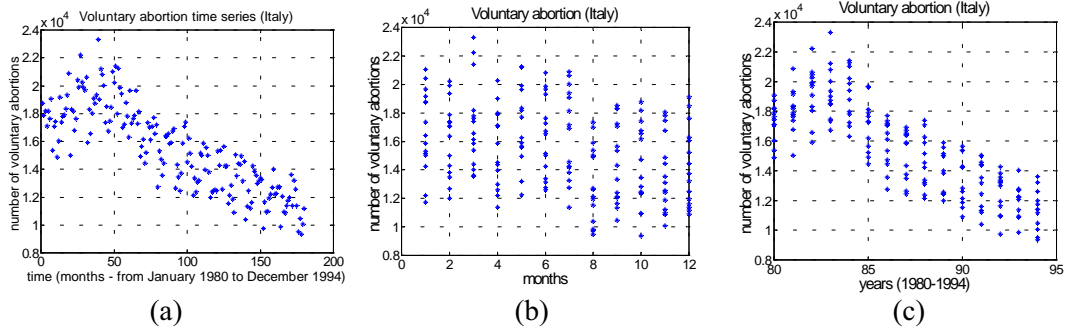


Figure 2: Monthly voluntary abortions in Italy from January 1980 to December 1994: (a) original time series; (b) data grouped by years; (c) data grouped by months

In Figures 2(b) and 2(c), the same data are represented grouped by years and by months, respectively. In Figure 2(b) only a qualitative increasing and decreasing trend is visible, whereas Figure 2(c) shows a slightly average decrease in the second part of the year, information that was also visible in the original VATS as depicted in Figure 2(a).

To separate trends and possible periodic components, we performed the Bayesian analysis described in detail in the previous section. After having performed 1 650 runs of the MCMC scheme on the VATS (1 500 plus 150 for the burn-in ⁽¹⁾), we were able to separate the trend and the periodic components, reported in Figures 3(a) and 3(b), with their 95 % confidence intervals.

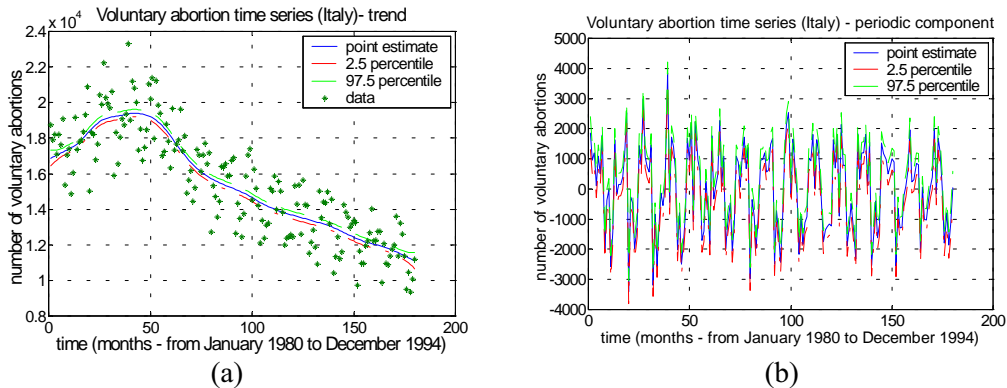


Figure 3. Monthly voluntary abortions in Italy from January 1980 to December 1994: (a) original data (stars) and reconstructed trend component with its 95 % confidence intervals; (b) the reconstructed periodic component with its 95 % confidence intervals

⁽¹⁾ The convergence of the Markov chain is verified by using the Raftery criterion (Raftery and Lewis, 1996). In particular, by choosing $q = \{0.025 \ 0.25 \ 0.5 \ 0.75 \ 0.975\}$, $r = \{0.02 \ 0.05 \ 0.01 \ 0.05 \ 0.02\}$ and $s = 0.95$, we verified that the burn-in is $M = 130$ and the required length of chain is $N = 1\,380$.

Unlike what one sees in the original data, the trend component is very regular; after an initial period in which voluntary abortions in Italy increased, starting in June–July 1983 voluntary abortions decreased constantly until the end of the study. Also, the periodic component shows an astonishing regularity in the studied period. Figure 4 shows a different view of the periodic component to highlight the presence of typical patterns by taking the point estimates of the cyclic component over the 15 years under study, superimposed on one another.

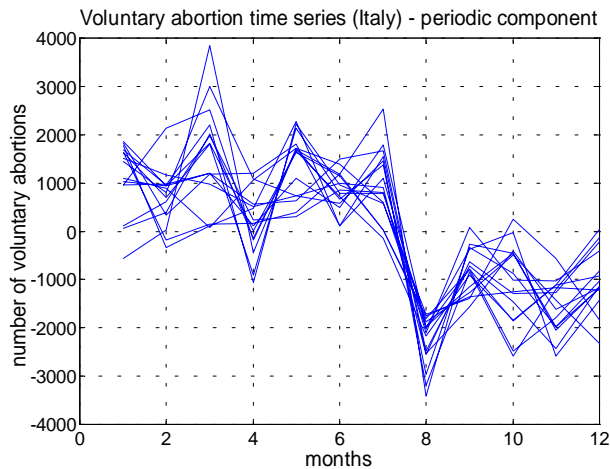


Figure 4: Voluntary abortions in Italy: the reconstructed (point estimates) periodic component for the years from 1980 to 1994

The periodic component clearly shows high values between January and June and low values in the other months with a well-defined minimum in August. Moreover, there is a clear decrease from March to April, and, finally, in the first part of the year (until June), even months seem to be locally lower than the odd months.

To better understand the results, we need to know if the highlighted behaviour of the VATS is caused by the dynamics of the births and by the seasonality of conceptions, which can be found by analysing the Italian birth time series (BTS).

4.1. The Italian birth time series

From the ISTAT publications, we have extracted the number of births in Italy for each month over the years 1980–95. We were unable to find data for 1992, but the Bayesian framework we are exploiting allowed us to handle this problem coherently.

The course of the BTS trend differs substantially from that of VATS. This suggests that the voluntary abortion dynamics are not completely explained from the course of births. Moreover, the birth periodic component is regular as well. In particular, the birth rate is roughly higher from May until October whereas it is lower in the other months. In addition, local differences are present among even and odd months, especially in the first half of the year.

To compare the dynamics of the BTS and VATS, we considered the voluntary abortion rate (VAR) defined in each month as the ratio between the number of abortions in that month and the total of conceived children. Formally,

$$VAR_i = \frac{VA_i}{VA_i + B_{i+6}} * 1000$$

The delay of six months between abortions and the correspondent births was chosen because abortions are generally performed during the third month of pregnancy.

4.2. The voluntary abortion rate

We can study how the voluntary abortion rate time series (VARTS) change in the years 1980–94. Because BTS data are not available for 1992, the VAR is not defined in the period from July 1991 to June 1992.

Performing the structural analysis on this time series in a way similar to the previous one (the only difference is the value of $\sigma_v^2 = 16$ and of $\gamma_1 = 10$), we can extract trend and cyclic components.

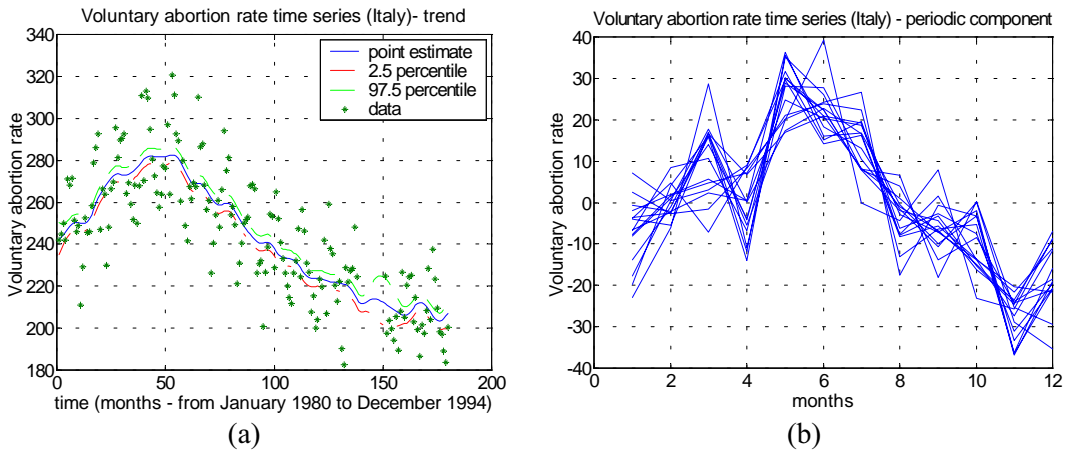


Figure 5: Monthly voluntary abortion rate in Italy from January 1980 to December 1994: (a) data (stars) and reconstructed trend component with its 95 % confidence intervals; (b) the reconstructed (point estimates) periodic component for the years from 1980 to 1994

Figures 5(a) and 5(b) show why births do not explain completely the dynamics of the VATS. In fact, the trend of VARTS clearly depicts that the number of abortions per 1 000 conceptions (approximated with the sum of births and voluntary abortions) changes over the time. In particular, after the approval of the abortion law, an increased number of VAR is apparent (from about 240/1 000 to about 280/1 000 in the 55th month — July 1984), whereas the number of abortions decreases constantly until the end of the considered period. Confidence intervals on the trend component are larger in 1991/92 (Figure 5(a), 138th–150th months) when no data are available.

The periodic component of VARTS is also not completely explained by the seasonality of the conceptions. In fact, the VARTS shows a significant periodic component (Figure 5(b)). Such a component is roughly constant over the studied period, and its amplitude (80) represents about 40–50 % of the mean value of the VARTS. Moreover, it is easy to see that the number of abortions over 1 000 conceptions is higher during spring months (March, May, June) and lower at the end of the year. April, August and November show relevant local changes.

5. Conclusions

In this paper, we have proposed a methodology to perform a structural Bayesian analysis on Official Statistics time series data.

1. The method can be effectively used in time series analysis, when data missing at random are present and when it is also important to derive interval estimates of trends or seasonality curves.
2. The method is particularly useful for detecting local changes in the structural components of the time series; such changes may then be related to external factors that should be hypothesised on the basis of the particular problem at hand. As a consequence, we believe that the computational machinery we used is more interesting for relatively short time horizon (or at least for a limited number of data) problems than for the analysis of massive data sets, collected over very long time periods or with very high frequency.
3. The Bayesian framework allows us to derive the distribution of all the parameters of interest given the data. Thus, from this distribution, it is possible to derive the point estimates and to quantify their credibility immediately.

For the application presented in this paper — the analysis of the voluntary abortions in Italy from January 1980 to December 1994 — the preliminary results suggest that this phenomenon is constantly decreasing and is also relative to the birth dynamics. Moreover, we found a surprising significant seasonality in voluntary abortions, also when data are compared with birth time series. This consideration can be an interesting starting point for a sociological analysis that might focus on temporal/regional differences and similarities in the voluntary abortions due, for example, to modification/difference in cultural backgrounds, habits or in the services provided by the national healthcare system.

The generality of the proposed method allows several extensions to our analysis. For example, by adopting a similar model, it is possible to derive a more parsimonious description of the time series assuming random fluctuations in the yearly components (in this way, the best trend and periodic components on the whole year are derived).

6. Acknowledgements

We thank ISTAT and, in particular, Dr Paolucci for having made available to us the voluntary abortion database. This work was supported by the European Union through the project BaKE (Esprit 29105).

7. References

- [1] Bellazzi, R., Magni, P. and De Nicolao, G. (2000), ‘Bayesian analysis of blood glucose time series from diabetes home monitoring’, *IEEE Transactions on Biomedical Engineering*, Vol. 47, pp. 971–975.
- [2] Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992), ‘A Monte Carlo approach to non-normal and non-linear state–space modelling’, *Journal of the American Statistical Association*, Vol. 87, pp. 493–500.
- [3] Carter, C. K. and Kohn, R. (1994), ‘On Gibbs sampling for state–space models’, *Biometrika*, Vol. 81, pp. 541–553.
- [4] Geman, S. and Geman, D. (1984), ‘Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, pp. 721–741.
- [5] Magni, P., Bellazzi, R. and De Nicolao, G. (1998), ‘Bayesian function learning using MCMC methods’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 1319–1331.
- [6] Raftery, A. E. and Lewis, S. M. (1996), ‘Implementing MCMC’, in Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov chain Monte Carlo in practice*, Chapman & Hall, London, pp. 115–130.
- [7] West, M. and Harrison, J. (1997), *Bayesian forecasting and dynamic models*, Springer, New York.

On the use of Bayesian networks to analyse survey data

Paola Sebastiani (*) and Marco Ramoni (**)

(*) *Department of Mathematics and Statistics, University of Massachusetts, United States*

(**) *Children's Hospital Informatics Program, Harvard University Medical School,
United States*

Keywords: Automated modelling, Bayesian networks, graphical models, Bayesian model selection

Abstract

This paper uses Bayesian modelling techniques to analyse a data set extracted from the British general household survey. The models used are Bayesian networks, which provide a compact and easy-to-interpret knowledge representation formalism. An issue considered is the need for automated Bayesian modelling.

1. Introduction

The general household survey is a yearly survey, based on a sample of the general population resident in private households in Great Britain. The general household survey began in 1971 and data is available from 1973 onwards. It is widely regarded as a 'gold standard' because of survey design and data collection and has been copied by several countries. The goal of this survey is to provide continuous information about the major social fields of population, housing, education, employment, health and income. Since the survey covers all these topics, it provides users with the opportunity to examine not only each topic separately, but also their mutual interplay. Summary of the statistical findings are published by the British Office of National Statistics, and are typically presented via contingency tables relating two or three variables at a time, (see Thomas et al., 1998). We believe that this communication style fails one of the primary objectives of the survey, which is to offer, to a non-technical audience, an up-to-date picture of living in Great Britain.

To avoid the fragmentation of the overall information, one should try to build a model that associates a large number of variables. To be a communication tool, however, such a model needs to be easily understandable, and easy to use. Understandability and usability being the requirements, we focus on Bayesian networks, which are known for providing a compact and easy-to-use representation of probabilistic information, (see Lauritzen, 1996, and Cowell et al., 1999). A Bayesian network has two components: a directed acyclic graph and a probability distribution. Nodes in the directed acyclic graph represent stochastic variables and arcs represent directed stochastic dependencies among these variables. Thus, the graph provides a simple summary of the dependency structure relating the variables. The probability distribution for the network variables decomposes according

to the conditional independencies represented by the directed acyclic graph, and each component — a conditional probability table — quantifies the remaining directed dependencies. The graph is an effective way to describe the overall dependency structure of a large number of variables, thus removing the limitation of examining the pair-wise associations of variables. Furthermore, one can easily investigate undirected relationships between the variables, as well as making a prediction and explanation, by querying the network. This last task consists of computing the conditional probability distribution of one variable, given that values of some variables in the network are observed. Nowadays there are several efficient algorithms for probabilistic reasoning, which take advantage of the network decomposability (Castillo et al., 1997), and commercial programs such as Bayesware Discoverer (available at <http://www.bayesware.com>) or Hugin (available at <http://www.hugin.com>) implement these algorithms.

The problem to be addressed, and we believe one of the reasons for the slow gain in popularity of these models in the statistical community, is how to practically build a Bayesian network from a large data set using Bayesian methods. This is considered in the next section. In Section 3 we analyse a data set extracted from the 1996 general household survey. The model selected is a network that displays a global picture of living in Britain and discovers interesting associations among variables describing the household wealth, the socioeconomic status and the ethnic group of the head of the household.

2. Overview of automated learning

A Bayesian network is a directed acyclic graph and a probability distribution. Nodes in the directed acyclic graph represent stochastic variables $X = (X_1, X_2, \dots, X_v)$, and directed arcs from parent nodes to a child node represent conditional dependencies. Any conditional dependence is quantified by the set of conditional distributions of the child variable given the configurations of the parent variables. Marginal and conditional independencies encoded by the directed acyclic graph (Lauritzen, 1996), provide the following factorisation of the joint probability distribution

$$p(x_{1k}, x_{2k}, \dots, x_{vk}) = \prod_{i=1}^v p(x_{ik} \mid \pi_{ij})$$

Here, $(x_{1k}, x_{2k}, \dots, x_{vk})$ is a combination of values of the variables in X . For each i , the variable Π_i denotes the parents of X_i while x_{ik} and π_{ij} denote the events $X_i = x_{ik}$, and $\Pi_i = \pi_{ij}$. Particularly, π_{ij} is the combination of values of the parent variable Π_i in the event $X = (x_{1k}, x_{2k}, \dots, x_{vk})$.

The problem we consider next is learning a Bayesian network from data. We can describe this as a hypotheses-testing problem. Suppose we have a set $M = \{M_1, M_2, \dots, M_g\}$ of

Bayesian networks, for the discrete random variables $X = (X_1, X_2, \dots, X_v)$. Each Bayesian network represents a hypothesis on the dependency structure relating the variables. We wish to choose one Bayesian network after observing a sample of data $D = \{x_{1k}, x_{2k}, \dots, x_{vk}\}$, for $k = 1, \dots, n$. With $p(M_h)$ denoting the prior probability of M_h , for each $h = 1, \dots, g$, the typical Bayesian solution to the model selection problem consists of choosing the network with maximum posterior probability

$$p(M_h | D) = \frac{p(D | M_h)p(M_h)}{p(M_h)}.$$

The quantity $p(D | M_h)$ is the marginal likelihood, and it is computed as follows. Given the Bayesian network M_h , let θ^h denote the vector parameterising the joint distribution of the variables $X = (X_1, X_2, \dots, X_v)$. We denote by $p(\theta^h)$ the prior density of θ^h . The likelihood function is $p(D | \theta^h)$ and the marginal likelihood is computed by averaging out θ^h from the likelihood function $p(D | \theta^h)$. Hence

$$p(D | M_h) = \int p(D | \theta^h)p(\theta^h)d\theta^h$$

The computation of the marginal likelihood requires the specification of a parameterisation of each model M_h , and the elicitation of a prior density for θ^h .

In this paper we suppose that the variables $X = (X_1, X_2, \dots, X_v)$ are all discrete, so that the parameter vector θ^h consists of the conditional probabilities $\theta^h_{ijk} = p(X_i = x_{ik} | \Pi_i = \pi_{ij}, \theta)$. In this framework, it is easy to show that, under the assumption of multinomial sampling with complete data, the likelihood function becomes

$$p(D | \theta^h) \propto \prod_{ijk} (\theta^h_{ijk})^{n_{ijk}}$$

where n_{ijk} is the sample frequency of pairs (x_{ik}, π_{ij}) in the database D . The Hyper-Dirichlet distribution, which is defined as a set of independent Dirichlet distributions $D(\alpha_{ij1}, \dots, \alpha_{ijc_i})$, one for each set of parameters $\{\theta^h_{ijk}\}_k$ associated with the conditional distribution $X_i | \pi_{ij}$, is a numerically convenient choice. It is well known (see Cowell et al., 1999), that this choice for the prior distribution provides the following formula for the marginal likelihood of the data:

$$p(D | M_h) = \prod_{ijk} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}.$$

Here, $n_{ij} = \sum_k n_{ijk}$ is the marginal frequency of π_{ij} in the database, and $\alpha_{ij} = \sum_k \alpha_{ijk}$.

For consistent model comparisons, we adopt symmetric Hyper-Dirichlet distributions, which depend on one hyperparameter α , called global precision. Each hyperparameter α_{ijk} is computed from α as $\alpha_{ijk} = \alpha / (q_i c_i)$, where c_i is the number of categories of the variable X_i , and q_i is the number of categories of the parent variable Π_i . The rationale behind this choice is to distribute the overall prior precision α in a uniform way among the parameters associated with different conditional probability tables. In this way, the prior probabilities quantifying each network are uniform, and all the prior marginal distributions of the network variables are uniform and have the same prior precision.

In principle, given a set of Bayesian networks, with prior probabilities, and a complete data set, one can compute their posterior probability distribution and select the network with maximum posterior probability. However, as the number of variables in the data set increases, the size of the search space makes the task infeasible. Thus some heuristic method is required to reduce the dimension of the search space. Fortunately, under some particular model prior probabilities, the posterior probability of each model M_h factorises, thus allowing local computations. This property can be fully exploited by imposing an order over the variables, which transforms model selection into a sequence of locally-exhaustive searches. We will also describe a greedy search algorithm to reduce the complexity of each locally-exhaustive search when the model space is still too large.

The marginal likelihood $p(D | M_h)$ above has a multiplicative form. This fact, together with the assumption that the network prior probabilities are decomposable (Heckerman et al., 1995), provides a factorisation of each model posterior probability. A prior probability for a network M_h is termed decomposable if it admits the factorisation

$$p(M_h) = \prod_{i=1}^v p(M_h^i),$$

where $p(M_h^i)$ is the prior probability of the local network structure that specifies the parent set Π_i for the variable X_i . Thus, decomposable priors are elicited by exploiting the modularity of a Bayesian network, and are based on the assumption that the prior probability of a local structure M_h^i of a Bayesian network is independent of the other parts M_h^j . This factorisation of each model prior probability, together with the factorisation of

the marginal likelihood, ensures that the posterior probability of the Bayesian network M_h can be written as

$$p(M_h | D) = \prod_{i=1}^v p(M_h^i | D) \propto \prod_{i=1}^v p(D | M_h^i) p(M_h^i).$$

Thus, the network posterior probabilities are decomposable and, in the comparison of models that differ only in the parent sets of a variable X_i , only the quantity

$$p(M_h^i | D) \propto p(D | M_h^i) p(M_h^i)$$

matters. Thus, for fixed i , the comparison of two local network structures M_h^i and \tilde{M}_h^i specifying different parent sets for X_i can be done by simply evaluating the product of the local Bayes factor

$$BF_{jl} = \frac{p(D | M_h^i)}{p(D | \tilde{M}_h^i)}$$

and the prior odds

$$\frac{p(M_h^i)}{p(\tilde{M}_h^i)}$$

to compute the posterior odds of M_h^i versus \tilde{M}_h^i . This comparison is independent of any other associations among the other $i - 1$ variables.

Now, the problem is how to exploit this posterior probability decomposability. One approach, proposed by Cooper and Herskovitz (see Cooper and Herskovitz, 1992), is to restrict the model search to a subset of all possible networks, which are consistent with an order relation \succ on the variables $X = (X_1, X_2, \dots, X_v)$. The order relation \succ is defined by $X_j \succ X_i$, if X_i cannot be a parent of X_j in any network in M . In other words, rather than exploring networks with arcs having all possible directions, this order limits the search to a subset of networks in which there are interesting directed associations.

At first glance, the requirement for an order among the variables appears to be a serious restriction on the applicability of this search strategy, but we have successfully implemented it in other applications. (see Sebastian et al, 2000) From a modelling point of view, specifying this order is equivalent to specifying the hypotheses to be tested and some careful screening of the variables in the data set may avoid the surprise of selecting a not

very sensible model or explore uninteresting associations. In the next section, we will consider the problem of selecting an order among the variables in a real application.

This order imposed on the variables, induces a set of k_i possible parents for each variable X_i , say $P_i = \{X_{i1}, X_{i2}, \dots, X_{ik_i}\}$. One way to proceed, which produces the sequence of locally-exhaustive searches, is to implement an independent model selection for each variable X_i as follows. For each variable X_i , we define M^i to be the set of networks given by the possible combinations of parents for X_i . The set of networks can be displayed on a lattice with k_i levels, each level having models in which the associated directed acyclic graph specifies k parents for X_i . The first level of the lattice contains the model M_0^i in which X_i does not have parents. The second level contains the k_i models M_j^i in which X_{ij} alone is parent of X_i and so on. For each variable X_i , the exhaustive search consists of evaluating the posterior probability of each model in the lattice so that the model with maximum posterior probability can be identified. The global model is then found by linking together the local models for each variable X_i .

Although the order among the variables greatly reduces the dimension of the search space, this locally-exhaustive search should explore a lattice of 2^{k_i} models for each variable X_i and, for large k_i , this may be infeasible. A further reduction is obtained via a greedy search strategy, also known as the K2 algorithm, (see Cooper and Herskovitz, 1992). The K2 algorithm is a bottom-up strategy, so that simpler models are evaluated first. For each variable X_i , rather than computing the posterior probability of all networks in the set M^i , the search moves up in the lattice as long as in the level just explored there is at least one network with posterior probability higher than the posterior probabilities of the networks in the precedent level. The search starts by evaluating the marginal likelihood $p(D | M_0^i)$ of the local network structure M_0^i encoding independence of X_i and the variables in the set P_i . The next step is the computation of the marginal likelihood $p(D | M_j^i)$ of the k_i Bayesian networks M_j^i , each of which describes the dependence of X_i on the variable X_{ij} . If the maximal marginal likelihood $p(D | M_j^i)$, for some J is greater than $p(D | M_0^i)$, the parent X_{iJ} is accepted and the search proceeds in the same manner by trying to add one of the parents from the set $P_i \setminus X_{iJ}$ to the Bayesian network selected. If none of the k_i Bayesian networks has a marginal likelihood greater than $p(D | M_0^i)$, the model M_0^i is accepted and the search moves to some other variable. Clearly, this heuristic search can end up in a local maximum, and one should be aware of this risk, when interpreting the model eventually selected. Other search strategies have been proposed to address this problem (see Cowell et al., 1999, and references therein).

3. Analysis

In this section, we analyse a data set extracted from the British general household survey ⁽²⁾, which was conducted between April 1996 and March 1997 by the Social Survey Division of the Office of National Statistics in the United Kingdom. This annual, multi-purpose survey is based on a sample of around 10 000 private households in Great Britain. Interviews are conducted with everyone aged over 16 in the household (around 18 000 adults). The data set we consider comprises 9 033 British households, which, following the definition introduced since 1981, consist of as a single person or of a group of people who have the address as their only or main residence and who share either one meal a day or the living accommodation.

In order to show the potential usefulness of our methodology, we selected 13 variables describing the British households in terms of composition (variables *Ad_fems*, *Ad_males*, *Children*, *Hoh_age*, *Hoh_gend*), regions of the United Kingdom (variable *Region*), one ethnicity indicator (variable *Hoh_origin*), one mobility indicator (variable *Hoh_reslen*) and economic indicators of the household (variables *Accom*, *Bedrms*, *Ncars*, *Hoh_status*, *Tenure*). A complete description of these variables and their states are summarised in Table 1. This group of variables was fully observed in the data set extracted from the survey.

The modelling of the data was carried out with the program Bayesware Discoverer, which implements the model search approach described in the previous section.

Table 1: Description of the variables extracted from the 1996 general household survey

Variable	Description	State description
<i>Region</i>	Region of birth of the Hoh	England, Scotland and Wales
<i>Ad_fems</i>	Number of adult females	0, 1, ≥ 2
<i>Ad_males</i>	Number of adult males	0, 1, ≥ 2
<i>Children</i>	Number of children	0, 1, 2, 3, ≥ 4
<i>Hoh_age</i>	Age of the Hoh	17-36; 36-50; 50-66; 66-98 (years)
<i>Hoh_gend</i>	Gender of the Hoh	Male, Female
<i>Hoh_origin</i>	Ethnic group of the Hoh	Cauc., Black, Chin., Indian, Other
<i>Hoh_reslen</i>	Length of residence	0-3; 3-9; 9-19; ≥ 19 (months)
<i>Hoh_status</i>	Status of Hoh	Active, Inactive, Retired
<i>Accom</i>	Type of accommodation	Room, Flat, House, Other
<i>Bedrms</i>	Number of bedrooms	1, 2, 3, ≥ 4
<i>Tenure</i>	House status	Rent, Owned, Social-Sector
<i>Ncars</i>	Number of cars	1, 2, 3, ≥ 4

⁽²⁾ Crown Copyright 1996. Used by permission of the British Office for National Statistics.

NB: Hoh denotes the head of the household — numbers of adult males, females and children refer to the household.

The approach to model selection described in the previous section requires the variables to be discrete. Therefore, the first step of the analysis was to discretise the continuous variables into four bins of approximately equal proportions. Before this step, variables having a skewed distribution were transformed in a logarithmic scale. Many integer-valued variables — as those indicating the number of adult males or females in the household — were appropriately recoded and States observed with low frequency were grouped into a unique state. We then choose the following order among the variables to limit the space of models to be explored.

$$\begin{aligned} &Region \succ Hoh_origin \succ Hoh_gend \succ Ad_fems \succ Ad_mal \succ Hoh_age \succ \\ &Hoh_status \succ Children \succ Tenure \succ Hoh_reslen \succ Accomod \succ Bedrms \succ Ncars. \end{aligned}$$

The choice was based on the following considerations. Geographic variables precede household variables and thus we are interested in conditioning on them first (e.g. see Thomas, et al., 1998). The ordering of some of the household demographic variables (e.g. *Hoh_origin*, *Hoh_gend*, *Ad_fems*, *Ad_males*, *Hoh_age*) and we chose the particular ordering for convenience. These variables are commonly thought of as explaining house wealth which is described by the variables *Hoh_status*, *Children*, *Tenure*, *Hoh_reslen*, *Accomod*, *Bedrms*, *Ncars*, while dependencies in which the age of the household head are directed influenced by any of these variables do not seem to be interesting. The remaining order was chosen in a similar way, on the basis of possible cause-effect relationships between the remaining variables.

We used this order to build four models, using the K2 algorithm, uniform prior probabilities on the possible networks, and symmetric Hyper-Dirichlet prior distributions for the model parameters. We chose four values for the global precision $\alpha = 1, 5, 10, 20$ to evaluate the effect of changing the global prior precision on the model selected. The evaluation was carried out by comparing the networks topologies, and their different predictive capabilities. This last aspect was evaluated by computing the classification accuracy of the four networks. Full details of the analysis are in Sebastiani and Ramoni (see Sebastiani and Ramoni, 2001) and led to select the network learned with $\alpha = 5$. This network is depicted in Figure 1 and is described in the next section.

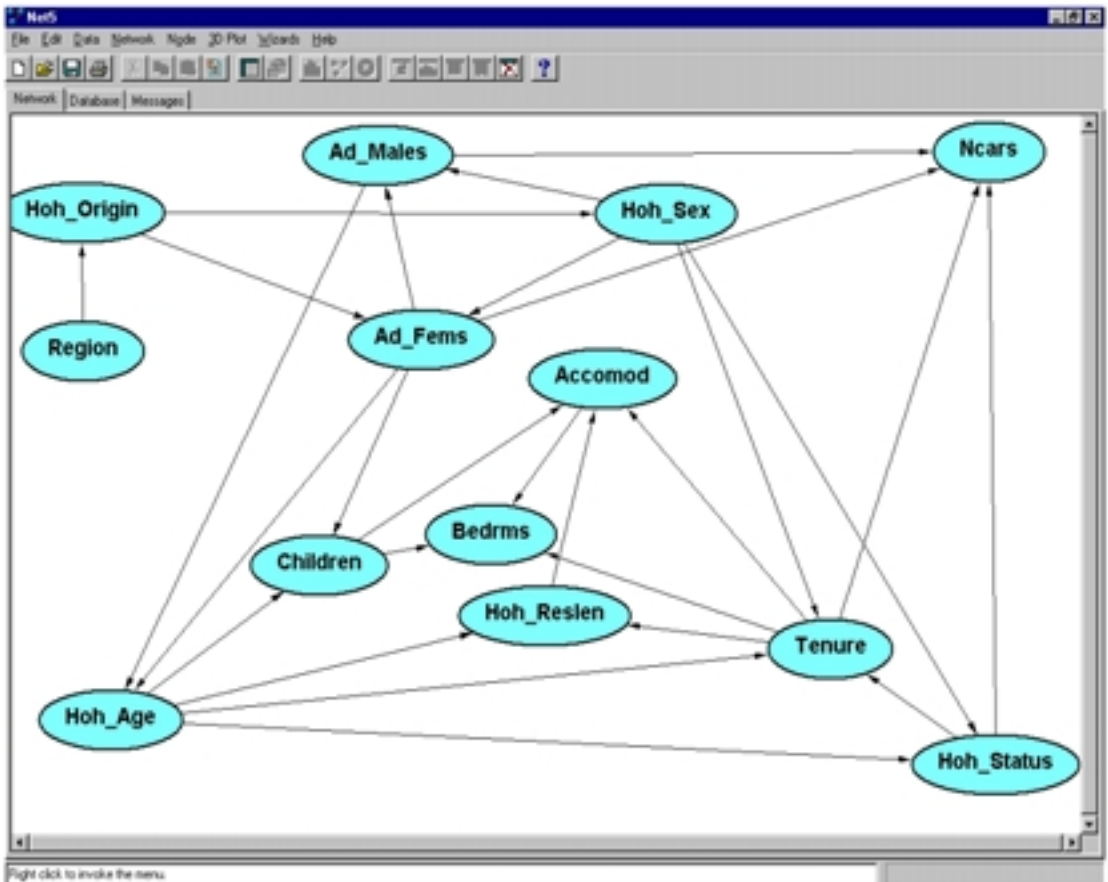


Figure 1: The Bayesian network selected from the data when the global prior precision α is 5

4. Results and discussion

The network in Figure 1 shows important, directed dependencies and conditional independencies. The dependency of the ethnic group of heads of the households on the variable *Region* reveals a more cosmopolitan society in England than Wales and Scotland, with a larger proportion of Blacks and Indians as head of households. The variables describing the ethnic group of the head of the household, of the gender of the head of the household, and the number of adult females in the household, separate *Region* from most of variables describing household wealth.

The working status of the head of the household (variable *Hoh_status*) is independent of the ethnic group given the gender and age of the head of the household. The estimated conditional probability table shows that when a young female is head of a household, she is much more likely to be inactive than a young male (40 % compared to 6 % when the age

group is 17–36). This difference attenuates as the age of the head of the household increases. The conditional distribution quantifying the dependency of the gender of the head of the household on the ethnic group reveals that Blacks have the smallest probability of having a male head of the household (64 %) while Indians have the largest probability (89 %).

The age of the head of the household depends directly on the number of adult males and females, and shows that households with no females and two or more males are more likely to be headed by a young male while, on the other hand, households with no males and two or more females are headed by a mid age female. There appear to be more single households headed by an elder female than an elder male. Furthermore, the composition of the household changes in the ethnic groups: the most interesting fact is that Indians have the smallest probability of living in a household with no adult males (10 %), while Blacks have the largest probability (32 %).

The tenure status of the accommodation depends directly on the age, gender and status of the household head. On average, the largest proportion of British households is established in owned accommodations (75 %), when the age of the head of the household is between 36 and 66 years. Younger heads of household have a larger chance of living in rented accommodations (20 %), while senior heads of household have a larger chance of living in accommodations provided by the social service (32 %). These figures however change dramatically when the gender of the head of the household is taken into account. When the head of the household is a young female, the probability that the household is in an owned accommodation is 27 %, against 65 % when the household head is a young male. This probability rises up to 52 % when the household head is an elder female compared to 69 % for elder males. Households are more likely to be in an accommodation provided by the social service when the head is an inactive female rather than an inactive male.

The number of bedrooms is directly affected by the number of children in the household, the type of accommodation and its tenure status. Households with two or more children are more likely to be in three bedroom flats or houses, but the accommodations provided by the social service are slightly smaller than those rented or owned by the head of the household. Houses are more likely to have a larger number of bedrooms than flats: the most likely number of bedrooms of an owned house is three, compared to one in a flat. Interestingly, flats provided by the social sector are more likely to be one-bed flats, while rented and owned flats are most likely to be two-beds flats. The length of residence is directly dependent on the age of the head of the household and the tenure status of the accommodation and shows that the length of residence in rented accommodations or those provided by the social service is shorter than that in owned accommodations.

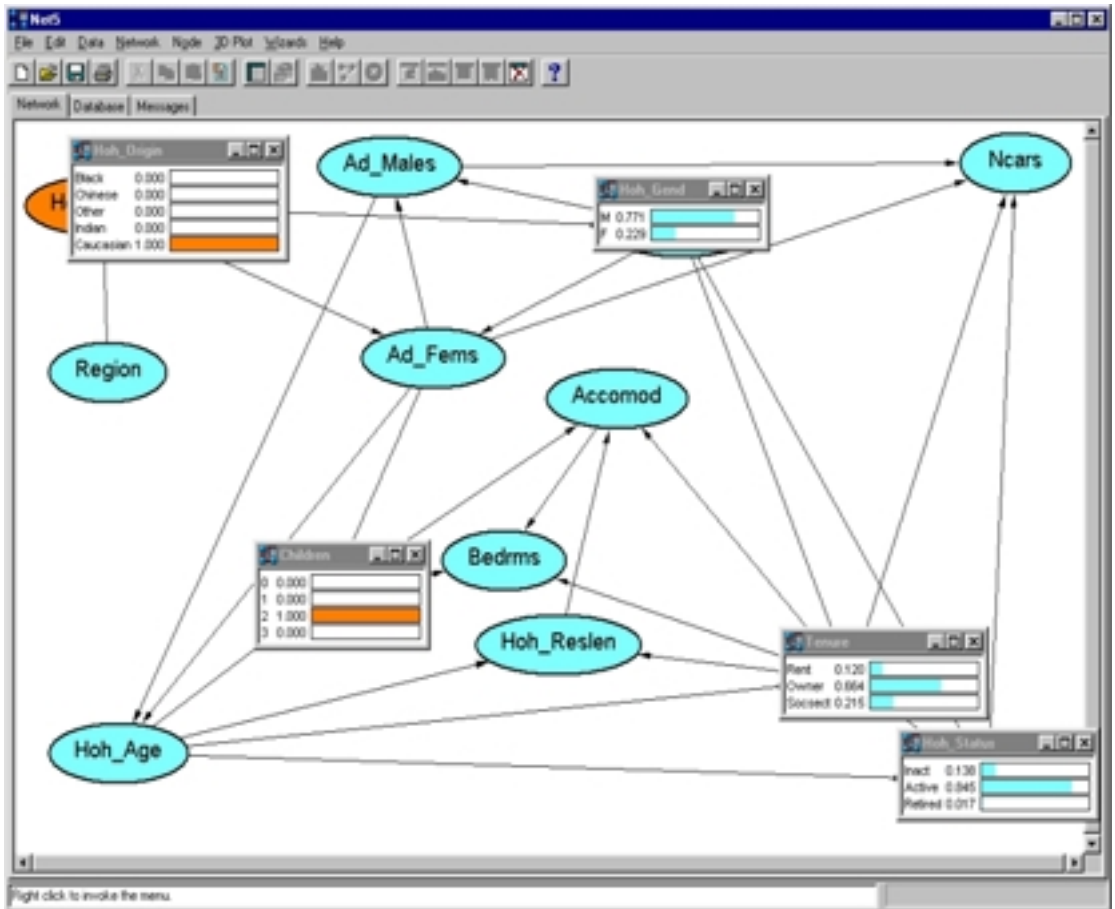


Figure 2: An example of query with the Bayesian network induced from the data

By querying the network, one may investigate other undirected associations and discover that, for example, the typical Caucasian mid family with two children has 77 % chances of being headed by a male who, with probability .57, is aged between 36 and 50 years. The probability that the head of the household is active is .84, and the probability that the household is in an owned house is .66. Results of these queries are displayed in Figure 2.

These figures are slightly different if the head of the household is, for example, Black. In this case, the probability that the head of the household is male (given that there are two children in the household) is only .62 and the probability that he is active is .79. If the head of the household is Indian, then the probability that he is male is .90 and the probability that he is active is .88. On average, the ethnic group changes slightly the probability of the household being in an accommodation provided by the social service (26 % for Blacks, 23 % for Chinese, 20 % Indians and 24 % Caucasians). Similarly, black heads of household are more likely to be inactive than heads of household from different ethnic groups (16 % Blacks, 10 % Indians, 14 % Caucasians and Chinese), and to be living in a less wealthy household, as shown by the larger probability of living in accommodations

with a smaller number of bedrooms and of having a smaller number of cars. Households headed by Blacks are less affluent than others, if the gender of the head of the household is not taken into account. However, the dependency structure shows that the gender of the head of the household and the number of adult females make all the other variables independent of the ethnic group. Thus, the model extracted suggests that differences in the household wealth are more likely caused by the different household composition, and in particular by the gender of the head of the household, rather than racial issues.

The robustness of many of these interpretations can be examined by careful alteration of the ordering of the variables and the structuring of the greedy search algorithm.

5. Conclusions

In this analysis, we focused on networks learned by using uniform model priors and sets of independent, symmetric Dirichlet distributions as prior distribution for each model parameters. The advantage of using these prior distributions is that they can be elicited simply by assigning the global prior precision and this choice produces consistent model comparisons. However, symmetric Dirichlet distributions are known to be too invariant, (see Forster and Smith, 1998), so that they model different dependency structures in the same way. One may wish to use a class of model parameter priors which encodes different prior information. An interesting challenge is to devise a class of prior distributions which maintains the consistency of model comparisons, feasibility of computations, and provides the user with more modelling freedom.

The analysis here was carried out by discretising all continuous variables, thus raising the issue of the effect of the discretisation. We are currently working on the implementation of a more general learning algorithm, which selects networks from data sets with both continuous and discrete variables.

One further issue is related to the publications of the results found with the method described here. A Bayesian network is not just the directed acyclic graph displaying the dependency structure selected, conditional on the data. It is also a probability distribution, and as such, the best way to publish the results is to give the entire network, and to let users make their own queries. Given the increasing importance that the world wide web is assuming nowadays as a communication system, publication of the network over the WWW offers a simple way to display results without giving direct access to the original data, thus preserving data confidentiality.

6. Acknowledgements

This research was supported by Eurostat, under contract EP29105. Material from the general household survey 1996 is Crown Copyright and has been made available by the Office for National Statistics through The Data Archive and has been used by permission. Neither the ONS nor The Data Archive bear any responsibility for the analysis or interpretation of the data reported here.

7. References

- [1] Castillo, E., Gutierrez, J. M. and Hadi, A. S. (1997), *Expert systems and probabilistic network models*, Springer, New York, NY.
- [2] Cooper, G. F. and Herskovitz, E. (1992), 'A Bayesian method for the induction of probabilistic networks from data', *Machine Learning*, Vol. 9, pp. 309–347.
- [3] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999), *Probabilistic networks and expert systems*, Springer, New York, NY.
- [4] Forster, J. J. and Smith, P. W. F. (1998), 'Model-based inference for categorical survey data subject to non-ignorable non-response (with discussion)', *Journal of the Royal Statistical Society, B*, Vol. 60, pp. 57–70.
- [5] Heckerman, D., Geiger, D. and Chickering, D. M. (1995), 'Learning Bayesian networks: The combinations of knowledge and statistical data', *Machine Learning*, Vol. 20, pp. 97–243.
- [6] Lauritzen, S. L. (1996), *Graphical models*, Oxford University Press, Oxford, UK.
- [7] Sebastiani, P. and Ramoni, M. (2001), 'Data analysis with Bayesian networks', under revision.
- [8] Sebastiani, P., Ramoni, M. and Crea, A. (2000), 'Profiling customers from in-house data', *ACM SIGKDD Explorations*, Vol. 1, pp. 91–96.
- [9] Thomas, M., Walker, A., Wilmot, A. and Bennet, N. (1998), *Living in Britain: Results from the 1996 general household survey*, The Stationery Office, London, UK.

Computing the posterior distribution of individual-level usual intakes with application to disease models

Michael J. Daniels and Alicia L. Carriquiry

Department of Statistics, Iowa State University, Ames, IA, United States

Keywords: attenuation, dietary surveys, Gibbs sampler, hierarchical models, informative priors, transformations, splines

Abstract

Diet has been linked to many health outcomes including breast cancer (Willett et al., 1992) and age-related maculopathy (ARM), a degenerative eye disease that may produce blindness. However, such relationships can be difficult to establish since the dietary quantity of interest, the usual intake of a nutrient, is unobservable in practice. The observable daily intake of a nutrient measures the usual intake with a large error. Furthermore, the distribution of this error is typically heterogeneous across individuals. In this paper, we present an approach to estimate the posterior distribution of individual-level usual intakes of a nutrient while accounting for heterogeneous measurement error in the intake of the nutrient. We then integrate over this distribution to obtain the posterior distribution of the regression coefficient associated with nutrient intake in a logistic regression model, adjusting for other risk factors. We illustrate the methods using dietary intake and ARM incidence data collected in the third national health and nutrition examination survey (NHANES III).

1. Introduction

The link between diet and certain health outcomes has been established (e.g. IOM, 2000). Investigators and practitioners alike agree that, most often, it is the chronic or long-term effect of diet on health that is of interest (NRC, 1986; Willett et al., 1992; Owusu et al., 1997). For example, it is not important if an individual's intake of vitamin C is very low on one day; what has been associated with poor health is the low habitual intake of vitamin C (e.g. NRC, 1986; Nusser et al., 1996). Let Y_{ij} denote the intake of the nutrient by individual i , $i = 1, \dots, n$ on day j , $j = 1, \dots, n_i$ and let w_i denote the usual nutrient intake by individual i . We define usual intake as the individual's long-run average intake of the nutrient, i.e.

$$w_i = E\{Y_{ij} \mid i\},$$

where the expectation is conditional on the individual. A simple additive measurement error model to represent the association between daily and usual intake for an individual was proposed by the NRC (1986). If e_{ij} is the measurement error associated with the i th

individual on the j th day, we get

$$Y_{ij} = w_i + e_{ij}, \quad (1)$$

where $w_i \sim (\mu_w, \sigma_w^2)$, $e_{ij} \sim (0, \sigma_{e,i}^2)$ and $\text{Cov}(w_i, e_{ij}) = 0$. Estimating the parameters in model (1) is not a straightforward matter; the error variances are heterogeneous across individuals, and the distributions of observed intakes Y_{ij} are typically very skewed, suggesting that the distributions of the unobservable usual intakes w_i may also be skewed.

Given the definition for w_i above, an unbiased estimate of w_i is \bar{Y}_i the mean of n_i days of observed daily intake for the individual. For large enough n_i , \bar{Y}_i is a reliable estimator of w_i , as \bar{Y}_i converges in distribution to w_i . Unfortunately, collecting dietary intake data is costly and fraught with errors. Several researchers have pointed to a strong time-in-sample effect, where individuals tend to under-report intake the longer they have been in the sample. Therefore, nationwide food consumption surveys such as the continuing survey of food intakes by individuals (CSFII; e.g. USDA, 1996) or the national health and nutrition examination survey (NHANES III; CDC, 1994) collect only one or two days of intake data for each individual in the sample. For those nutrients that are not widely available in the food supply, such as most antioxidants, the within-individual variance in intakes is significantly larger than the between-individual variance in the population (e.g. Sempos et al., 1985; Nusser et al., 1996; Carriquiry, 1999).

In this paper, we address the problem of estimating the regression coefficient associated with usual intake of a nutrient when incidence of age-related maculopathy (ARM) is the categorical response. ARM is a degenerative eye disease that may lead to blindness. Presence of the disease is modelled as a function of the usual intake of a nutrient and of a vector of other covariates a_i measured for the i th individual in the sample. We first describe an approach for approximating the marginal posterior distributions, $\pi_i(w_i | \mathbf{Y})$, of usual intake for each individual in the sample. Given these marginal posterior distributions, we then propose a method for incorporating uncertainty about the w_i when fitting a model $\text{Pr}(\text{ARM}) = f(w, a; \beta)$. We illustrate the method by applying it to data collected in NHANES III (CDC, 1994).

The paper is organised as follows. In Section 2, we describe the intake and health outcome data collected in NHANES III. The model and estimation of model parameters are discussed in Section 3. The application of the methods described is shown in Section 4. We conclude with brief comments and suggestions for future work in Section 5.

2. Intake and health outcome data

The United States has been conducting nationwide food consumption surveys since 1936. Currently, the two major food intake surveys are the CSFII, administered by the US Department of Agriculture, and the NHANES, which is conducted by the National Centre for Health Statistics of the Centres for Disease Control and Prevention. The latest CDC survey is NHANES III, which was in the field from 1988 to 1994. Both surveys obtain information on socio-demographic variables and food intake for all sampled individuals. In addition, the NHANES also collects extensive information on health status and outcomes.

The survey instrument used to collect food intake data is called a 24-hour recall form, and, as the name suggests, respondents are asked to recall what they ate during the past 24 hours. In the NHANES, a complete medical examination (including medical history, X-rays, and blood tests) is carried out as well. The total sample size in NHANES III is approximately 30 000 individuals. Low-income individuals, African-Americans and Hispanics were oversampled, so post-survey weights are available to ‘correct’ inferences to the general population. Overall, only about 5 % of respondents in NHANES III were asked to provide a second day of intake data. Thus, information for estimating the usual intake of nutrients is scarce.

We used information on food intake, socio-demographic factors, and ARM incidence for males, aged 50 years of age or older, collected in NHANES III. We focus on estimating the relationship between usual intake of vitamin C and the incidence of ARM.

Daily food intakes were collected for 1 933 males aged 50 years and older. Food intakes were converted into nutrient intakes using ‘maps’ from the food conversion databases constructed and maintained by the US Department of Agriculture. We considered only nutrient intakes from food sources; since over 30 % of the US population consume at least one dietary supplement, the nutrient intakes we used for our analyses are very likely to underestimate the actual nutrient intakes by individuals in the sample.

The presence of early or late ARM was measured on all individuals in the sample, using gradable retinal fundus photography as the diagnostic tool. A standard protocol (see, for example, Goldberg et al., 1988) was then used to classify individuals into one of three categories: no ARM, early ARM, and late ARM. The protocol requires investigation of the following factors: presence of any drusen, retinal pigment epithelial depigmentation, retinal pigmentation, geographic atrophy, and signs of exudative macular degeneration. For the purpose of our analyses, we aggregated the early ARM and late ARM categories into one, and considered a bivariate response variable with two levels: healthy individuals and individuals affected with ARM. Of the 1 933 men with complete dietary intake information, 1 833 had complete covariate information (smoking, ethnic group, eye colour) and reliable ARM diagnosis. The healthy individuals numbered 1 544, and those diagnosed with ARM numbered 279.

A partial description of the data used for analysis is given in Table 1. In the table, we present the proportion of individuals afflicted with ARM in the presence and absence of factors known to increase the chance of suffering from ARM: blue-eyes, smoking, and Caucasian ethnic group. We also show, for healthy and ARM patients, the observed mean daily intake of vitamin C (in mg) and the standard error of the mean.

Table 1: The upper portion of the table shows the proportion of men diagnosed with ARM in the presence (Yes) or absence (No) of risk factors — The lower table gives the mean and the standard error (SE) of the mean of continuous risk factors in ARM patients and in healthy individuals

	Yes	No
Blue eyes	86/525=16 %	193/1 308=15 %
Smoking	246/1 582=16 %	47/351=13 %
White (race)	199/1 100=18 %	94/833=11 %
	Mean (SE), ARM patients	Mean (se), healthy individuals
Age	74 (0.49)	70 (0.18)
Vitamin C	91 (4.3)	92 (2.2)

The distribution of observed vitamin C intakes among the 1 933 men in the sample is skewed, with a long tail to the right. Furthermore, the within-individual standard deviations in vitamin C intake are positively associated with individual mean intakes; the more the individual consumes of the nutrient, the more variable the intake of the nutrient from one day to the next.

Because of the small number of individuals in the sample for whom a replicate day of intake data was collected, we decided to use informative prior distributions for some of the parameters in the model described in the next section. The informative prior distributions we constructed were based on a similar nationwide food consumption survey, the 1994–96 CSFII, administered by the US Department of Agriculture. The CSFII survey uses a survey instrument that is similar to the one used in NHANES III. The surveys are not strictly comparable, however, as in CSFII some of the 24-hour recalls are administered by telephone. Furthermore, the design of the samples are different, and the demographic groups that are targeted by each survey are also different. CSFII however, collects two days of intake data for each individual in the sample. Clearly, this sample information can be incorporated into the NHANES III analysis via appropriate prior distributions, to supplement the reduced information about the measurement error variance. We discuss the choice of informative prior distributions in more detail in Subsection 3.3.

3. Model and estimation of model parameters

3.1. Logistic regression model

Let D_i , $i = 1, \dots, n$, be a binary disease indicator for the i th individual, and let w_i be a dietary risk factor. Further, let \mathbf{a}_i be a vector of other risk factors. We write

$$D_i | w_i, \mathbf{a}_i \sim \text{Bernoulli}(p_i),$$

where

$$p_i = \frac{\exp\{\beta_0 + \beta_1 w_i + \sum_{l=2}^L \beta_l a_{i,l}\}}{1 + \exp\{\beta_0 + \beta_1 w_i + \sum_{l=2}^L \beta_l a_{i,l}\}}.$$

Here, $a_{i,l}$, $l = 2, \dots, L$ is the value of the l th covariate measured on the i th individual, and β_0 is the usual regression intercept. We are interested in the estimation of β_l .

Assume that we have measured a_i without error, but that w_i is measured with error. It is well known (e.g. Carroll et al., 1984; Rosner et al., 1992) that if the measurement error of w_i is not adjusted for, the estimate of w_i will be *attenuated*, that is, $|\hat{\beta}_l|$ will be too small. We propose a Bayesian approach for the estimation of β_l that takes into account the uncertainty about the value of the usual intake w . Throughout, we choose a flat prior distribution for β . In Section 3.2, we describe an approach to estimate the posterior distribution $\pi_i(w_i | \mathbf{Y})$, of the **true** usual intake w_i for the i th individual. We are ultimately interested in $\pi(\beta | \alpha, D, \mathbf{Y})$, the marginal posterior distribution of the vector of parameters β , so we will need to evaluate the integral in expression (2) once the $\pi_i(w_i | \mathbf{Y})$ are available. An expression for the posterior distribution of interest is

$$\pi(\beta | \mathbf{a}, D, \mathbf{Y}) \propto \int \prod_i \pi(D_i | \beta, w_i, \mathbf{a}_i) \pi_i(w_i | \mathbf{Y}) dw_i. \quad (2)$$

We proceed in two steps. We first develop an approach to approximate the individual marginal posterior distributions of usual nutrient intake $\pi_i(w_i | \mathbf{Y})$ (as described in Section 3.2). We then implement a Monte Carlo integration algorithm to evaluate (2) (see Section 3.3).

3.2. Estimating individual posterior distributions of usual intake

The method we use is similar to the one given in Daniels and Carriquiry (1999). In general, daily intakes Y_{ij} are not normally distributed. Thus, we transform the Y_{ij} into normal random variables, X_{ij} . So that this transformation is flexible enough to accommodate any dietary component, we developed a semi-parametric transformation approach that is carried out in two steps. In the first step, the best (in the minimum mean squared error sense) power transformation γ to the normal scale is estimated. In the second step, we fit a natural cubic spline to map power-transformed intake observations into the normal space. Consider the pairs (Y_{ij}^γ, z_{ij}) , where the z_{ij} are the corresponding normal scores. The model we postulate is

$$Y_{ij}^\gamma = \alpha_0 + \sum_{p=1}^3 \alpha_p z_{ij}^p + \sum_{p=1}^k \alpha_{p+3} t_p + \varepsilon_{ij}, \quad (3)$$

where $t_p = (z_{ij} - r_p)^3 I_{z_{ij} \geq r_p}$ and the ε_{ij} are normal random variables with mean 0. The number of join points k , the location of the join points r_1, r_2, \dots, r_k and the parameters $(\alpha_0, \alpha_1, \dots, \alpha_{k+4})$ are unknown. Because the Y_{ij}^γ are sample quantiles, the residuals ε_{ij} are not independent. We specify the variance of the residuals to be proportional to the asymptotic covariance matrix of the sample quantiles, Y_{ij}^γ , V (for the specific form of V , see Schervish, 1996, p. 404-410); i.e., $\text{Var}(\varepsilon) = \sigma^2 V$. We let $g(Y_{ij}) = X_{ij}$ denote the transformation function.

To fit the cubic spline to our data, we proceed in a Bayesian fashion and use the approach of Denison et al. (1998). A reversible jump MCMC sampler (Green, 1995) was implemented for computations. To facilitate the computations, we diagonalise V . Daniels and Carriquiry (1999) demonstrate that this has little effect on the posterior distribution of usual intakes. From the sequence of transformations thus generated, we draw a sample of size m_I of transformation functions. We then create m_I samples of size $N = \sum_{i=1}^n n_i$ each from the original non-normal observations. For more details on the transformation model and its implementation, see Daniels and Carriquiry (1999).

For each transformed sample, we formulate a hierarchical measurement error model with three levels. In *level 1*, the individual's transformed daily intake, X_{ij} is modelled as a normally distributed random variable with mean equal to x_i , the usual intake of the nutrient for subject i , and with a subject-specific measurement error variance:

$$X_{ij} | x_i, \sigma_{ui}^2 \sim N(x_i, \sigma_{ui}^2).$$

In *level 2*, we model the heterogeneity in the usual intakes and in the measurement error variances across individuals:

$$\begin{aligned} x_i | \mu_x, \sigma_x^2 &\sim N(\mu_x, \sigma_x^2), \\ \log(\sigma_{ui}^2) | \mu_A, \sigma_A^2 &\sim N(\log(\mu_A), \sigma_A^2). \end{aligned} \quad (4)$$

Finally, in *level 3* we place flat priors on the hyper-parameters (μ_x, σ_x^2) . Since the sample contains little information about the subject-specific measurement error variances, σ_{ui}^2 , we construct informative priors for the hyperparameters (μ_A, σ_A^2) . Details are given in the next section.

Draws of all parameters in the model can be generated using a Gibbs sampler with a Metropolis step for sampling values of the measurement error variances. For each of the m_1 samples of transformed daily intakes, we generate m_2 draws $\{x_1, \dots, x_n, \sigma_{u1}^2, \dots, \sigma_{un}^2, \mu_x, \sigma_x^2, \mu_A, \sigma_A^2\}$.

For any of the m_1 samples, the x_i in level 2 of the measurement error model represent the true individual usual intakes in the transformed scale. The final step back-transforms the x_i draws into the original scale. We denote the back-transformed values as w_i , and by definition,

$$w = E\{Y | x = x_k\} = E\{g^{-1}(x + u) | x = x_k\}, \quad (5)$$

where the subscript k indexes the k th draw of x from its marginal posterior distribution and u is the measurement error. To approximate the expectation in (5), we generate, for each draw $\{x_{ik}, \sigma_{uik}^2\}, k = 1, \dots, m_2$, a large number q of measurement errors u_{ij} from their distribution $N(0, \sigma_{uik}^2)$, and then compute the Monte Carlo mean. The backtransformed w_i (one for each individual in each of the $m_1 \times m_2$ replicates) is a draw from the posterior distribution of the usual intake for individual i , $\pi_i(w_i | \mathbf{Y})$. Thus, using this procedure, we obtain a sample from the posterior distribution of usual intake for each subject in the sample.

Once we have obtained a sample from the posterior distribution of usual intakes for the subjects in the study, as detailed above, we sample from the posterior distribution $\pi(\beta | \mathbf{a}, D, \mathbf{Y})$ by numerically evaluating the integral given in expression (2). Specifically, for each draw from $\pi_i(w_i | \mathbf{Y})$, we generate S draws from the conditional posterior distribution of β (conditional on w_i), using a random walk Metropolis-Hastings algorithm with a multivariate normal candidate distribution with mean 0 and covariance

matrix proportional to the inverse of the observed information matrix for the maximum likelihood estimator of β , conditional on the current sample from the posterior distribution of w_i .

3.3. Construction of informative prior distributions

About 8 % of males aged 50 and older in NHANES III were asked to provide a second day of dietary intake information. Clearly, the sample information about the parameters of the measurement error distribution is very sparse. As a result, we specified informative prior distributions for the parameters of the within individual measurement error distribution, μ_A and σ_A^2 . To do so, we use nutrient intake data collected in a similar nationwide food consumption survey, the CSFII 1994–96, which included two observations of each of the nutrient intake of 1 636 males 50 years of age and older.

To derive values for the hyperparameters from the CSFII data, we applied the methods in this paper to the CSFII data using non-informative priors wherever possible, and obtained the posterior distributions of the measurement error model parameters. We used these estimated posterior distributions to specify the hyperparameters μ_A and σ_A^2 in level 3 of model (4). For simplicity, we chose conjugate priors: $\log(\mu_A)$ was a priori normally distributed, and $1/\sigma_A^2$ was distributed as a gamma random variable. For the former, the values of the parameters were obtained by matching the moments of the prior to the posterior distribution of $\log(\mu_A)$ from the CSFII data. For the latter, we parameterised the Gamma distribution as $G(\delta m, \delta)$ where the first moment, m , was matched to the posterior mean obtained from the CSFII data for $1/\sigma_A^2$. The parameter δ can be thought of a 'sample size' that reflects our confidence in the value chosen for m .

4. Results

We applied the methodology presented above to the data described in Section 2. The posterior distributions of usual intake for six individuals in our sample are given in Figure 1; note that the posterior distributions of the usual intakes for different individuals are highly variable. For example, the individual in the upper left panel has a low usual intake of vitamin C, and there is considerable uncertainty about the usual intake of the individual shown in the upper middle panel.

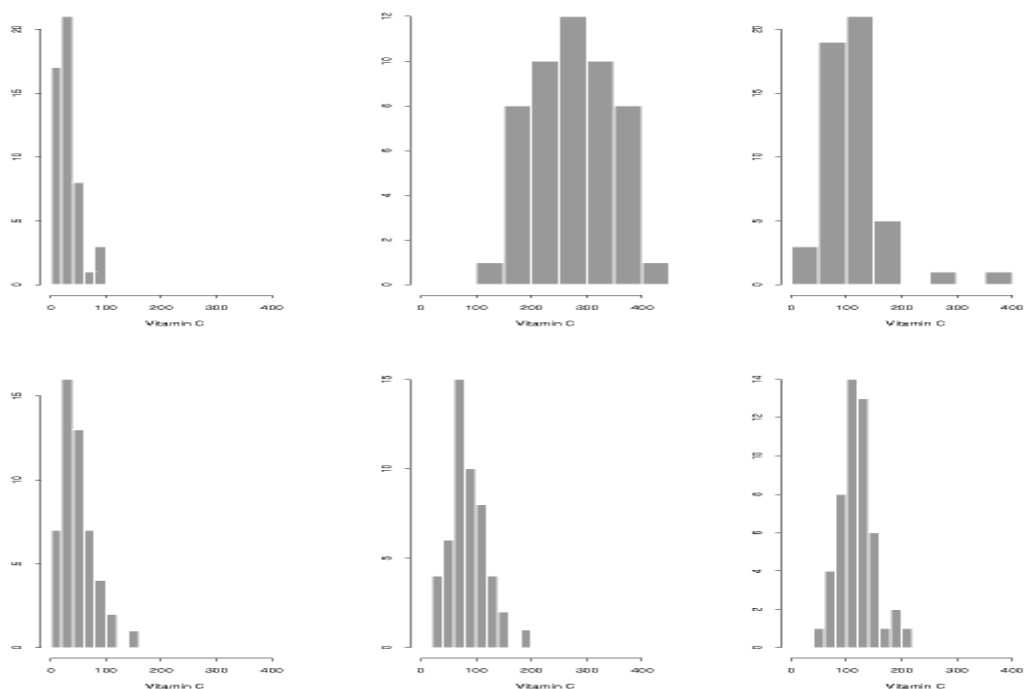


Figure 1: Marginal posterior distribution of the usual intake for six subjects

Table 2 shows the mean and the standard deviation of the posterior distribution of the regression coefficients associated with risk factors other than diet. A 95 % credible interval is also shown. Note that individuals are more likely to develop ARM when they are older, smoke, are Caucasian, and have blue eyes, as expected. The association between smoking and ARM incidence is not strong in this data set, however. This may be due to measurement error in the recording of smoking status.

Table 2: Posterior means, standard deviations, and 95 % credible intervals of the regression coefficients associated with covariates other than vitamin C intake

	Mean (standard deviation)	95 % credible interval
Intercept	7.10 (.76)	(5.6, 8.6)
Age	.069 (.009)	(.052,.088)
Smoking	.15 (.19)	(-.24,.52)
Race (white)	.43 (.16)	(.11,.75)
Eye colour (blue)	.28 (.16)	(-.04,.59)

We estimated β_I using the approach we propose and also two other alternatives. The two alternative methods are the following:

- Instead of integrating out w_i as in (2), plug in the observed mean intake for each individual. The observed mean intake measures usual intake with error. The approximation to the posterior distribution of β is then proportional to $\prod_i \pi(D_i | \beta, \mathbf{a}_i, \bar{Y}_i)$.
- Instead of the observed mean intake, we plugged in the posterior mean of the distribution $\pi_i(w_i | \mathbf{Y})$. In this case, the approximation to the posterior distribution of β was proportional to $\prod_i \pi(D_i | \beta, \mathbf{a}_i, E[w_i | \mathbf{Y}])$.

Estimates of the mean and the standard deviation of the marginal posterior distribution of β_I , the regression coefficient associated with vitamin C intake, are given in Table 3 for each of three estimation approaches. The corresponding 95 % credible sets are listed as well.

Table 3: Posterior means, standard deviations, and 95 % credible intervals for $\beta_I \times 1000$, using three different approaches

Method	Mean (standard deviation)	95 % credible interval
Plug in \bar{Y}	.00 (.80)	(-1.5, 1.6)
Plug in $E(w_i \mathbf{Y})$	-.53 (1.6)	(-3.6, 2.6)
Average over $\pi(w_i \mathbf{Y})$	-.19 (1.3)	(-2.3, 2.5)

We did not detect a significant association between vitamin C intake and incidence of ARM. In all three approaches, the regression coefficient β_I was not significantly different from zero. As expected, however, less attenuation of β_I , and thus larger estimated effects, was observed for the approaches that adjusted for measurement error.

The negative sign of the regression coefficient associated with vitamin C intake was unexpected. Further exploration of the data, however, revealed that vitamin C intake appears to be higher in those individuals diagnosed with late ARM than in those diagnosed with early ARM. This suggests that a serious confounding effect is operating; individuals diagnosed with ARM prior to the time of examination in NHANES III seem to have changed their diet as a response to the diagnosis.

To assess how much information about the within-individual measurement error parameters was available in the NHANES sample, we compared the prior and posterior distributions of the measurement error parameters $\log(\mu_A)$ and $1/\sigma_A^2$. The prior distribution and the sample-updated posterior distribution of $\log(\mu_A)$ were rather similar (see Figure 2), indicating that the sample did not contribute much information about the value of the central moment of $\log(\sigma_{ui}^2)$. The prior and posterior mean of $1/\sigma_A^2$ were essentially the same, but the posterior variance was larger than the prior variance.

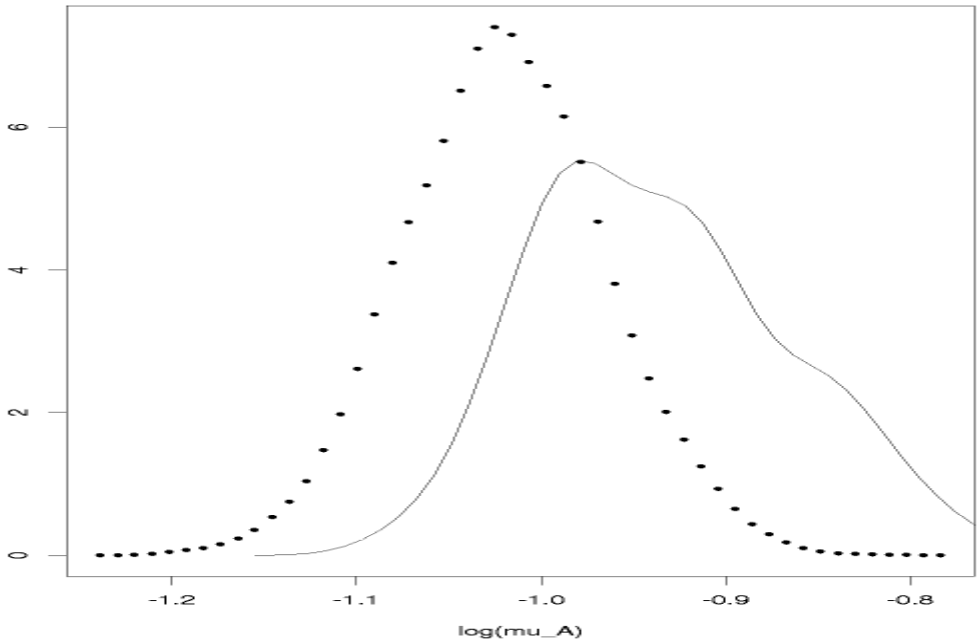


Figure 2: Prior and posterior for $\log(\mu_A)$ — the dotted line denotes the prior distribution and the solid line the posterior distribution

5. Discussion

We have proposed an approach for estimating individual posterior distributions of usual intake of a dietary component that may be measured with heterogeneous (across individuals) error. If the nutrient in question is a dietary risk factor for the onset of a disease, then values of usual intake of the component for each individual can be sampled from the posterior distribution and used as regressors in a logistic (or ordinary) regression model. We intend to further explore the sensitivity of inferences on β to varying specifications of informative prior distributions on the measurement error parameters. We also plan to conduct simulation studies to assess the performance of the proposed approach

under correct and incorrect assumptions about the measurement error distribution. A recent paper by DiMatteo et al. (2000) addresses several problems with the Denison et al. (1998) approach to curve fitting; we are currently implementing such modifications (although we believe that they will have little impact on our ultimate inferences).

Clearly, intakes of more than one nutrient typically need to be considered in association with disease. We are currently working on extending the methodology to allow for multiple covariates measured with error. This will involve approximating multivariate distributions of usual intakes, since intakes of different nutrients are typically correlated.

Finally, given the characteristics of dietary survey data, we realise that a naive regression model like the one we have formulated cannot capture the effect of some serious potential confounders. For example, individuals that are classified during the NHANES III examination as suffering from late ARM, may have been initially diagnosed with the disease years before, and may have changed their diet as a result of the diagnosis. That may explain the incorrect sign we obtained for the regression coefficient associated with vitamin C intakes.

6. References

- [1] Carriquiry, A. L. (1999), ‘Assessing the prevalence of nutrient inadequacy’, *Public Health Nutrition*, Vol. 2, pp. 23–33.
- [2] Carroll, R. J., Spiegelman, C. H., Lan K. K. G, Bailey, K. T. and Abbott, R. D. (1984), ‘On errors-in-variables for binary regression models’, *Biometrika*, Vol. 71, pp. 19–25.
- [3] (Centres for Disease Control and Prevention, National Centre for Health Statistics)(1994), *Plan and operation of the third national health and nutrition examination survey, 1988-94*, Hyattsville, Maryland, DHHS Publication No (PHS) 94—1308.
- [4] Daniels, M. J. and Carriquiry, A. L. (1999), ‘Dietary assessment and estimation of intake densities’, in Paulino, C. D, Pacheco, A. and Ferreira da Cunha, A. P (eds), *Afirmar a Estatística: Um Desafio para o Século XXI*, Sociedade Portuguesa de Estatística, Lisbon, Portugal.
- [5] Denison, D. G. T., Mallik, B. K. and Smith, A. F. M. (1998), ‘Automatic Bayesian curve fitting’, *Appl. Statist.*, Vol. 60, pp. 333–350.
- [6] DiMatteo, I., Genovese, C. R. and Kass, R. E. (2000), ‘Bayesian curve fitting with free-knot splines’, *Carnegie Mellon Department of Statistics, Technical Report*, No 716.

- [7] Goldberg, J., Flowerdew, G., Smith, E., Brody, J. and Tso, M. O. M. (1988), 'Factors associated with age-related macular degeneration', *American Journal of Epidemiology*, Vol. 128, pp. 700–710.
- [8] Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika*, Vol. 82, pp. 711–732.
- [9] NRC (National Research Council)(1986), *Nutrient adequacy: assessment using food consumption surveys*, National Academy Press, Washington, DC.
- [10] Nusser, S. M., Carriquiry, A. L., Dodd, K. W. and Fuller, W. A. (1996), 'A semi-parametric transformation approach to estimating usual daily intake distributions', *Journal of the American Statistical Association*, Vol. 91, pp. 1440–1449.
- [11] Owusu, W., Willett, W. C., Feskanih, D., Ascherio, A., Spiegelman, D. and Colditz, G. A. (1997), 'Calcium intake and the incidence of forearm and hip fractures among men', *Journal of Nutrition*, Vol. 127, pp. 1782–1787.
- [12] Rosner, B., Spiegelman, D. and Willet, W. C. (1992), 'Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error', *American Journal of Epidemiology*, Vol. 136, pp. 1400–1413.
- [13] Schervish, M. (1996), *Theory of statistics*, Springer-Verlag, New York.
- [14] Sempos, C. T., Johson, N. E., Smith, E. L. and Gilligan, C. (1985), 'Effects of intraindividual and interindividual variation in repeated dietary records', *American Journal of Epidemiology*, Vol. 121, pp. 120–130.
- [15] USDA (United States Department of Agriculture) Agricultural Research Service (1991), *Continuing survey of food intakes by individuals*, 1989–91, CSFII Report No 91-4, US Government Printing Office, Washington, DC.
- [16] USDA (United States Department of Agriculture) Agricultural Research Service (1996), *Continuing survey of food intakes by individuals*, 1994–96, CSFII Report No 96-3, US Government Printing Office, Washington, DC.
- [17] Willett, W. C., Hunter, D. J., Stampfer, M. J., Colditz, G. A., Manson, J., Spiegelman, D., Rosner, B., Hennekens, C. H. and Speizer, F. E. (1992), 'Dietary fat and fibre in relation to risk of breast cancer', *Journal of the American Medical Association*, Vol. 286, pp. 2037–2044.

Semi-parametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries

Ngianga B. Kandala, Stefan Lang, Stephan Klasen and Ludwig Fahrmeir

University of Munich, Ludwigstraße 33, Germany

Keywords: developing countries, semi-parametric Bayesian inference, spatial models, undernutrition

Abstract

We estimate semi-parametric regression models of chronic undernutrition (stunting) using the 1992 demographic and health surveys (DHS) from Tanzania and Zambia. We focus particularly on the influence of the child's age, the mother's body mass index, and spatial influences on chronic undernutrition. Conventional parametric regression models are not flexible enough to cope with possibly non-linear effects of the continuous covariates and cannot flexibly model spatial influences. We present a Bayesian semi-parametric analysis of the effects of these two covariates on chronic undernutrition. Moreover, we investigate spatial determinants of undernutrition in these two countries. Compared to previous work with a simple fixed effects approach for the influence of provinces, we model small-scale district specific effects using flexible spatial priors. Inference is fully Bayesian and uses recent Markov chain Monte Carlo techniques.

1. Introduction

Acute and chronic undernutrition is considered to be one of the worst health problems in developing countries. As one of the most important indicators of deprivation, undernutrition is of intrinsic concern to policy-makers. In addition, it is also associated with other important development outcomes such as high mortality and poor labour productivity (Sen, 1999; Unicef, 1998). In fact, some estimates claim that undernutrition is implicated in over 50 % of deaths in developing countries (Unicef, 1998).

Undernutrition among children is usually determined by assessing the anthropometric status of the child relative to a reference standard. Researchers distinguish between three types of undernutrition: wasting or insufficient weight for height indicating acute undernutrition; stunting or insufficient height for age indicating chronic undernutrition; and underweight or insufficient weight for age which could be a result of either. Wasting, stunting, and underweight for a child i are typically determined using a Z-score which is defined as:

$$Z_i = \frac{AI_i - MAI}{\sigma}$$

where AI refers to the individual's anthropometric indicator (e.g. height at a certain age), MAI refers to the median of the reference population, and σ refers to the standard deviation of the reference population. The reference standard typically used for the calculation is the NCHS-CDC growth standard that has been recommended for international use by WHO (WHO, 1983, 1995).

Important determinants of undernutrition include the education, income, and nutritional situation of the parents, access to clean water and sanitation, and primary healthcare, and immunisation facilities (Unicef, 1998; Klasen, 1999; Nyovani et al., 1999). Some of these influences are likely to have non-linear effects on undernutrition. In particular, the impact of the nutritional situation of the parents, measured using the body mass index (BMI , defined as the weight in kilograms divided by the square of height in metres) on the child's nutritional status is presumed to follow an inverse U-shape. Parents who exhibit a very low BMI , indicating their poor nourishment, are likely to have poorly nourished children. At the same time, parents with a very high BMI might also have poorly nourished children as the obesity associated with their high BMI indicates poor quality of nutrition and might therefore indicate poor quality of nutrition for their children.

Moreover, the development of undernutrition typically follows a pattern that is closely related to the age of the child. While some children are already born undernourished due to growth retardation *in utero*, the anthropometric status of children worsens considerably only after four to six months, when children are weaned and solid foods are introduced (WHO, 1995; Stephenson, 1999). This is due to the influence of poor quality nutrition that is replacing breast milk as well as the onset of infectious diseases. These diseases are often related to unclean water and food which is replacing the breast milk, and the child no longer profits from the mother's antibodies that were transmitted through the breast milk (Stephenson, 1999). Initially, the worsening anthropometric status shows up as acute undernutrition. But then stunting develops and worsens until about the age of two to three. At that time, the body has, through reduced growth, adjusted to reduced nutritional intake and now needs relatively fewer nutrients to maintain this smaller stature. In addition, the body has developed its immune system to fight the impact of infectious diseases more effectively (WHO, 1995; Moradi and Klasen, 2000).

Even after controlling for the impact from these well-known correlates, researchers have found important spatial differences in indicators such as undernutrition, or mortality in many developing countries (World Bank, 1995). They may be related to left-out variables that have a distinct spatial pattern. Obvious examples of such variables are the disease environment in certain areas, the climate which may affect the quality of nutrition and the persistence of illness, access to important infrastructure (such as health centres, major roads or railroads), regional economic opportunities and constraints, etc. (Gallup and Sachs, 1998). To the extent that undernutrition is directly affected by the presence or absence of infectious diseases, such spatial patterns may also capture the spatial distribution of certain infectious epidemics.

In this paper, we model the determinants of stunting (i.e. chronic undernutrition) in Zambia and Tanzania. Stunting rates are high in both countries. Overall, 42 % of Zambian

children under age five are classified as stunted (Z score less than minus 2) and 18 % as severely stunted (Z score less than minus 3). In Tanzania, some 43 % of children under five are classified as stunted and 18 % are severely stunted (Somerfelt and Stewart, 1994).

A particular focus of our analysis is the use of a flexible approach to model the impact of the child's age and the mother's BMI on undernutrition as well as consider spatial effects with the help of a semi-parametric Bayesian modelling approach developed by Fahrmeir and Lang (2001a,b) and Lang and Brezger (2001). In a related paper (Kandala et al., 2001), spatial effects have been included by using simple fixed effects for provinces. In the current paper spatial random effects models are used to determine small-scale regional (district-specific) effects. The results give refined insight into spatial effects on undernutrition. Inference is fully Bayesian and uses recent Markov chain Monte Carlo (MCMC) techniques.

2. Semi-parametric Bayesian regression models

2.1. Observation model

Consider the regression situation, where observations (y_i, x_i, w_i) , $i=1, \dots, n$, on a metrical response y , a vector $x = (x_1, \dots, x_p)$ of metrical covariates and a vector $w = (w_1, \dots, w_r)$ of categorical covariates are given. We assume that y_i given the covariates and unknown parameters are independent and Gaussian with mean η_i and a common variance σ^2 across subjects, i.e. $y_i \sim N(\eta_i, \sigma^2)$. In our application on childhood undernutrition the response is stunting measured as a Z-score. Traditionally, the effect of the covariates on the response is modelled by a linear predictor

$$\eta_i = x_i' \beta + w_i' \gamma. \quad (1)$$

In this paper particular emphasis is on the effects of the two metrical covariates 'age of the child' *AGC* and the 'mother's body mass index' *BMI*, which are possibly non-linear, and on regional effects of the district where the mother and child live. Thus, we replace the strictly linear predictor (1) by the more flexible semi-parametric predictor

$$\eta_i = f_1(x_{i1}) + \dots + f_p(x_{ip}) + f_{spat}(s_i) + w_i' \gamma.$$

Here f_1, \dots, f_p are non-linear smooth effects of the metrical covariates and f_{spat} is the effect of district $s_i \in \{1, \dots, S\}$ where mother i lives. In a further step we may split up the spatial effect f_{spat} into a spatially correlated (structured) and an uncorrelated (unstructured) effect

$$f_{spat}(s_i) = f_{str}(s_i) + f_{unstr}(s_i).$$

A rationale is that a spatial effect is usually a surrogate of many unobserved influences, some of them may obey a strong spatial structure and others may be present only locally. By estimating a structured and an unstructured effect we aim at separating between the two

kinds of factors. As a side effect we are able to assess to some extent the amount of spatial dependency in the data by observing which one of the two effects is larger. If the unstructured effect exceeds the structured effect, the spatial dependency is smaller and vice versa. Such models are common in spatial epidemiology, see e.g. Besag et al. (1991).

2.2. Prior assumptions

In a Bayesian approach unknown functions $f_j, j=1, \dots, p, f_{str}, f_{unstr}$ and parameters γ as well as the variance parameter σ^2 are considered as random variables and have to be supplemented with appropriate prior assumptions. In the absence of any prior knowledge we assume independent diffuse priors $\gamma_j \propto \text{const}, j=1, \dots, r$, for the parameters of fixed effects. Another common choice are highly dispersed Gaussian priors.

Several alternatives are available for the priors of the unknown (smooth) functions f_1, \dots, f_p . For the moment we may distinguish roughly two main approaches for Bayesian semi-parametric modelling. These are base functions approaches with adaptive knot selection (e.g. Denison et al., 1998, Biller, 2000, and Smith and Kohn, 1996) and approaches based on smoothness priors. In the following we will focus on the latter one. Several alternatives have been proposed for specifying a smoothness prior for the effect f of a metrical covariate x . Among others, these are random walk priors (Fahrmeir and Lang, 2001a, Kandala, Lang and Klasen 2001), Bayesian smoothing splines (Hastie and Tibshirani, 2000) and Bayesian P-splines (Lang and Brezger, 2001). In this paper we focus on P-splines.

The basic assumption behind the P-splines approach is that an unknown smooth function f of a particular covariate x can be approximated by a spline of degree l defined on a set of equally-spaced knots $\zeta_0 = x_{\min} < \zeta_1 < \dots < \zeta_{r-1} < \zeta_r = x_{\max}$ within the domain of x . It is well known that such a spline can be written in terms of a linear combination of $m = r+l$ B-spline basis functions B_t , i.e.

$$f(x) = \sum_{t=1}^m \beta_t B_t(x),$$

The basis functions B_t are defined locally in the sense that they are nonzero only on a domain spanned by $2+l$ knots. It would be beyond the scope of this paper to go into the details of B-splines and their properties, see e.g. de Boor (1978). The vector $\beta = (\beta_1, \dots, \beta_m)$ is unknown and must be estimated from the data. In a simple regression spline approach the unknown regression coefficients are estimated using standard methods for fixed-effects parameters. However, a crucial point with simple regression splines is the choice of the number and the position of knots. For a small number of knots the resulting spline space may be not flexible enough to capture the variability of the data. For a large number of knots estimated curves may tend to overfit the data. As a remedy to these problems Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of

adjacent regression coefficients to guarantee sufficient smoothness of the fitted curves. In a Bayesian approach, we replace difference penalties by their stochastic analogues, i.e. first or second order random walk models for the regression coefficients

$$\beta_t = \beta_{t-1} + u_t, \quad \beta_t = 2\beta_{t-1} - \beta_{t-2} + u_t,$$

with Gaussian errors $u_t \sim N(0, \tau^2)$ and diffuse priors $\beta_1 \propto \text{const}$, or β_1 and $\beta_2 \propto \text{const}$, for initial values, respectively. A first order random walk penalises abrupt jumps $\beta_t - \beta_{t-1}$ between successive states and a second order random walk penalises deviations from the linear trend $2\beta_{t-1} - \beta_{t-2}$. Random walk priors may be equivalently defined in a more symmetric form by specifying the conditional distributions of parameters β_t given its left and right neighbours, e.g. β_{t-1} and β_{t+1} in the case of a first order random walk. Then, random walk priors may be interpreted in terms of locally polynomial fits. A first order random walk corresponds to a locally linear and a second order random walk to a locally quadratic fit to the nearest neighbours, see e.g. Besag et al. (1995). The amount of smoothness is controlled by the additional variance parameter τ^2 , which corresponds to the smoothing parameter in a frequentist approach. The larger (smaller) the variance, the rougher (smoother) are the estimated functions.

Let us now turn our attention to the spatial effects f_{str} and f_{unstr} . For the spatially correlated effect $f_{str}(s)$, $s=1, \dots, S$, we choose Markov random field priors common in spatial statistics (Besag, et al. 1991). These priors reflect spatial neighbourhood relationships. For geographical data one usually assumes that two sites or regions s and r are neighbours if they share a common boundary. Then a spatial extension of random walk models leads to the conditional, spatially autoregressive specification

$$f_{str}(s) | f_{str}(r), r \neq s \sim N\left(\sum_{r \in \partial_s} f_{str}(r) / N_s, \tau^2 / N_s\right)$$

where N_s is the number of adjacent regions, and $r \in \partial_s$ denotes that region r is a neighbour of region s . Thus the (conditional) mean of $f_{str}(s)$ is an average of function evaluations $f_{str}(s)$ of neighbouring regions. Again the variance τ^2_{str} controls the degree of smoothness.

For a spatially uncorrelated (unstructured) effect f_{unstr} a common assumption is that the parameters $f_{unstr}(s)$ are i.i.d. Gaussian

$$f_{unstr}(s) | \tau^2_{unstr} \sim N(0, \tau^2_{unstr}).$$

For a fully Bayesian analysis, variance or smoothness parameters τ^2_j , $j=1, \dots, p$, str , $unstr$, are also considered as unknown and estimated simultaneously with corresponding unknown functions f_j . Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by highly dispersed inverse gamma distributions $p(\tau^2_j) \sim IG(a_j, b_j)$ with known hyperparameters a_j and b_j .

2.3. Posterior inference

Bayesian inference is based on the posterior and is carried out using recent MCMC simulation techniques. Let α denote the vector of all unknown parameters in the model. Then, under usual conditional independence assumptions, the posterior is given by

$$P(\alpha | y) \propto \prod_{i=1}^n L_i(y_i, \eta_i) \prod_{j=1}^p \{p(\beta_j | \tau_j^2) p(\tau_j^2)\} p(f_{str} | \tau_{str}^2) p(f_{unstr} | \tau_{unstr}^2) \prod_{j=1}^r p(\gamma_j) p(\sigma^2)$$

where $\beta_j, j=1, \dots, p$, are the vectors of regression coefficients corresponding to the functions f_j . The full conditionals for the parameter vectors β_1, \dots, β_p as well as the full conditionals for f_{str}, f_{unstr} and fixed effects parameters γ are multivariate Gaussian. For the variance components $\tau_j^2, j=1, \dots, p, str, unstr$, and σ^2 the full conditionals are inverse gamma distributions. Thus, a Gibbs sampler can be used for MCMC simulation, drawing successively from the full conditionals for $\beta_1, \dots, \beta_p, f_{str}, f_{unstr}, \tau_j^2, j=1, \dots, p, str, unstr$, and σ^2 . Efficient sampling from the Gaussian full conditionals of non-linear functions is guaranteed by Cholesky decompositions for band matrices. More details can be found in Rue (2001), Fahrmeir and Lang (2001b) and Lang and Brezger (2001).

3. Data and results

The Demographic Health Surveys (DHS) of Tanzania and Zambia, both conducted in 1992, are used in this study. These surveys are produced jointly by Macro International, a USAID-funded firm specialising in demographic research, and the national statistical agency of the respective country. They draw a representative sample of women of reproductive age and then administer a questionnaire and an anthropometric assessment of themselves and their children that were born within the previous five years. The data set contains information on family planning, maternal and child health, child survival, HIV-AIDS, educational attainment, and household composition and characteristics. There are 8 138 cases for Tanzania and 6 299 for Zambia. The sample is drawn through stratified clustered sampling and draws, in the case of Zambia, 262 clusters from 53 districts in Zambia. In the case of Tanzania, we have data from 357 clusters drawn from 25 regions (which, to make them compatible with Zambia, we refer to as districts for our analysis). These districts can be grouped into nine provinces in Zambia and six provinces in Tanzania.

One cannot assume that the clusters selected in each district are fully representative of the districts in which they are located, as the surveys only attempted to generate a fully representative sample at the provincial level. Consequently, the spatial analysis will be affected by some random fluctuations. Some of this random variation can be reduced through the structured spatial effects as it includes neighbouring observations in the analysis. It should, however, be pointed out that such a spatial analysis should preferably be applied to census data, the most important official demographic data source in most

developing countries, where the precision of the spatial analysis would be much higher. Unfortunately, most censuses do not collect data on undernutrition and often the full dataset is not available for such analyses.

We concentrate in the analysis on the flexible modelling of the effects of the child's age, the mother's *BMI* and the districts on chronic undernutrition (stunting), measured using the Z-score as described above, which we standardised to take on a mean of 0 and a standard deviation of 1. In addition, we consider several categorical variables including the sex of the child, the education and employment situation of the mother, access to water (later omitted as it was found to have a negligible influence) and locality (urban and rural). All categorical variables are included in effect-coding (rather than as usual dummy variables) and in the tables we also report on the reference category. The education variable is coded in three categories called, respectively, 'no education and incomplete primary education' (reference category), 'complete primary education and incomplete secondary education', and 'complete secondary education and higher'. For the employment situation of the mother, we distinguish between working and not working. We estimate separate models for each country with predictor

$$\eta = \gamma_0 + f_1(AGC) + f_2(BMI) + f_{spat}(s) + \gamma'w$$

where w includes the categorical covariates in effect coding. The functions f_1 and f_2 are modelled by cubic P-splines with second order random walk penalty. For the spatial effect f_{spat} we experimented with different prior assumptions. For both countries we estimated models where either a structured or an unstructured effect was included as well as a model where both effects were included. Based on these results we found clear evidence for both countries of spatial correlation among neighbouring districts. Hence, a spatially correlated effect f_{str} is included in the predictors of our final models. For Zambia, we additionally include an unstructured effect f_{unstr} because there is evidence of local extra variation in the highly urbanised areas of Zambia. For Tanzania an unstructured effect is excluded from the final model. All computations have been carried out with *BayesX*, a software package for Bayesian inference based on MCMC simulation techniques, see Lang and Brezger (2000).

Table 1 shows the results of the fixed-effects parameters in Tanzania. Despite modelling the spatial effects differently here, the results for the (non-spatial) fixed effects are virtually identical to Kandala, Lang, and Klasen (2001).

The substantive findings are generally as expected. Children of highly-educated mothers living in urban areas are better nourished than other children. Children of working mothers do slightly worse. Being female is also associated with reduced levels of stunting, a finding consistent with Svedberg (1996) and Klasen (1996).

The results are quite similar for Zambia (Table 2). The direction of influences are the same in both countries. The size of the coefficients differ slightly. In particular, both the effect of education and of residence (urban versus rural) is somewhat smaller in Zambia.

Moreover, the 80 % credible region for the mother's employment status now includes zero. Access to water was found to be insignificant in both countries and was therefore omitted.

Figures 1 to 4 show the non-linear effects of child's age and the mother's *BMI*. Also here, the differences to Kandala, Lang, and Klasen (2001) which was based on a different prior are very minor. Moreover, the results are not sensitive to the additional inclusion of non-linear regional effects, suggesting that the method applied is able to separately identify non-linear covariate and regional effects. Figure 1 shows the effect of the *BMI* of the mother in Tanzania. Shown are the posterior means together with 80 % pointwise credible intervals. As hypothesised, we find the influence to be in the form of an inverse U shape. While the inverse U looks nearly symmetric, the descending portion exhibits a much larger range in the credible region. This appears quite reasonable as obesity of the mother (possibly due to a poor quality diet) is likely to pose less of a risk for the nutritional status of the child as very low BMIs which suggest acute undernutrition of the mother. The Z-score is highest (and thus stunting lowest) at a *BMI* of around 30-35.

Figure 2 shows the effect of the child's age on its nutritional status in Tanzania. As suggested by the nutritional literature, we are able to discern the continuous worsening of the nutritional status up until about 20 months of age. This deterioration sets in right after birth and continues, more or less linearly, until 20 months. Such an immediate deterioration in nutritional status is not as expected as the literature typically suggests that the worsening is associated with weaning at around four to six months. One reason for this unexpected finding could be that, according to the surveys, most parents give their children liquids other than breast milk already shortly after birth which might contribute to infections at these early ages.

After 20 months, stunting stabilises at a low level. Through reduced growth and the waning impact of infections, children are apparently able to reach a low-level equilibrium that allows their nutritional status to stabilise.

We also see a slight improvement of the Z-score around 24 months of age. This is picking up the effect of a change in the data set that makes up the reference standard. Until 24 months, the currently-used international reference standard is based on white children in the United States of high socioeconomic status, while after 24 months, it is based on a representative sample of all US children (WHO, 1995). Since the latter sample exhibits worse nutritional status, comparing the Tanzanian children to that sample leads to a sudden improvement of their nutritional status at 24 months. This anomaly of the reference standard is one reason for WHO's current efforts to construct a new reference standard (WHO, 1999).

Figure 3 shows the effect of mother's *BMI* on chronic undernutrition in Zambia. Also here we find a, somewhat less pronounced, inverse U-shape. The inverse U-shape is much more pronounced on the ascending left portion than on the descending right portion, which is only barely discernible. Again, this is consistent with the notion that acute undernutrition of the mother is more of a risk for the child than obesity. Figure 4 shows the impact of the child's age on stunting in Zambia. Here the deterioration in the nutritional status appears to be slightly longer. It only stabilises at around 22-24 months. Also here, we see a slight

improvement in the Z-score around 24 months associated with the change in the reference population.

To explore district-specific spatial effects, Figures 5-11 explore the spatial effects of undernutrition in the two countries. As mentioned above, in Tanzania we report on the model that only includes structured effects, while in the case of Zambia we report on the model that includes both structured and unstructured effects. Figure 5 shows the structured random effects for Tanzania and Figure 6 indicates the significance of the observed spatial effects in the form of a posterior probability map. The levels correspond to significantly negative (black coloured), significantly positive (white coloured) and insignificant (grey coloured). Two important observations emerge. First, there is a strong south north gradient in these regional effects with a fairly sharp dividing line running through the centre of the country. Over and above the impact of the fixed effects, there appear to be negative influences on undernutrition in the south that are quite general and affect most of the regions there. Given that the southern districts all are at significantly lower elevation than the rest of the country, it is likely that climatic and associated disease factors are responsible for this pronounced regional pattern. Second, living in the capital Dar es Salaam is associated with significantly better nutrition despite being surrounded by areas with negative regional effects on undernutrition. Living in the capital must thus provide access to nutrition and healthcare that is superior in ways that have not been captured adequately in the fixed effects.

To compare our district-specific non-linear effects with our simple fixed effects for provinces which we used in Kandala et al. (2001), Figure 7 presents a map that shows those provincial effects for the six provinces. Note that one can only distinguish five provinces as the effects for the Central province and the Coastal province are virtually identical. These crude provincial fixed effects miss most of the findings we discussed above. In particular, the sharp south–north divide present in the district analysis is now no longer visible as the Central and Coastal provinces include districts on both sides of that divide. Moreover, the positive effect of Dar es Salaam is simply averaged in with the Coastal province. Clearly, a lot is lost when relying on this crude strategy of modelling spatial effects.

Figures 8 to 10 show the structured and the unstructured random effects for Zambia. The structured effects show a sizeable difference between significantly worse undernutrition in the northern parts of the country (in particular the districts in Luapula and Northern province), and significantly better nutrition in the central and south-western parts. These regional patterns are similar, but not identical to analyses of poverty and deprivation undertaken by the World Bank (World Bank, 1995). In terms of income poverty, the World Bank found poverty to be lowest in the central parts of the country. In addition, poverty was also much lower along the main trunk road and railroad lines even outside the central part of the country. In terms of deprivation (based on a mean score of various service items), the World Bank also found Luapula province among the worst off, while it surprisingly included the Central province and the North-western province among the worst-off regions. While we also find Luapula province to be among the worst off in the

country, our analysis shows a clearer geographic pattern with the north-east being worst off and the central and south-western districts being best off.

The unstructured random effects are mostly not significant. But they nevertheless point in interesting directions. In particular, they suggest a fair amount of variation over and above the structured effects. Particularly noteworthy is the fact that for some urban centres, the unstructured effects point to higher undernutrition, once the fixed effects (which include a positive effect of urban areas) and the structured effects are controlled for. This is particularly noteworthy for Kitwe in the Copperbelt, but also visible for Lusaka and Kabwe in the central part of the country. In contrast to Tanzania, it thus appears that some urban agglomerations are associated with worse nutrition. This may be related to the impact of economic decline and adjustment policies which have hit the Copperbelt and some other urban areas particularly hard (World Bank, 1995).

Figure 11 shows the provincial fixed effects used in Kandala et al. (2001). While the overall spatial structure is more or less accurately reproduced, the effects of urban agglomerations on the structured and unstructured effects distort the picture particularly for the Copperbelt and the Central province where most of these urban agglomerations are.

In sum, the flexible modelling of the district-specific effects paints a much more nuanced picture than was presented by the regional fixed effects and thus gives a better impression of the spatial variation of undernutrition. Moreover, the semi-parametric Bayesian approach used is able to identify subtle influences of the mother's *BMI*, the child's age on the nutritional status of the child.

These findings are not only relevant for analytical purposes but have considerable policy significance. In particular, the age effect points to considerable nutritional problems immediately after birth, possibly related to the use of unclean liquids. This is a subject that should be investigated further. Second, the non-linear influence of the *BMI* indicates that not only parental undernutrition, but parental malnutrition might also have negative effects on the nutritional status of children. Third, the regional influences on undernutrition also are of great policy significance. In particular, they suggest that, in Tanzania inhabitants of the capital are much less affected by undernutrition, even if they suffer similar risk factors (as captured by the fixed effects). The same is, however, not true in Zambia, where some urban agglomerations are associated with higher undernutrition. Also, more emphasis must be placed upon the role of remoteness as well as climatic and geographic factors on undernutrition. The south-north divide in Tanzania and the regional effects in Zambia bear out the importance of such considerations.

4. Conclusion

In this paper, we have applied a semi-parametric Bayesian approach to model the determinants of chronic undernutrition (stunting) in Tanzania and Zambia. The fixed effects show the importance of the mother's education, employment status, residence, and

the sex of the child on chronic undernutrition. We also find that our methods are identifying subtle effects of the mother's *BMI*, the child's age, and regional influences on undernutrition. In particular, the effects of the *BMI* on the child's nutritional status appear to be in the form of an inverse U. Moreover, stunting appears to worsen until about 20–25 months and then stabilises at a low-level equilibrium. Furthermore, we find sizeable regional effects. In both countries, we are able to pick up a distinct regional pattern of undernutrition that is not adequately captured by relying on provincial fixed effects.

Given the limitations of spatial analysis when the database is a household survey, an important message emerging from this research is that it would be very worthwhile to have census data and other official data sources to undertake such detailed spatial analyses. With such data sources, much more detailed and more precise spatial structures could be uncovered which would be highly relevant for both analytical as well as policy purposes.

5. Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 'Statistische Analyse diskreter Strukturen'. We thank Macro International for providing the data and the precise locations of the clusters sampled.

6. Tables

Table 1: Fixed effects for Tanzania (effect coding)

Variable	Mean	10 % quantile	90 % quantile
Constant	0.29	0.17	0.41
Working	– 0.02	– 0.04	0
Not working	0.02	0	0.04
No education and incomplete primary education	– 0.26	– 0.35	– 0.17
Complete primary education and incomplete secondary education	– 0.18	– 0.26	– 0.09
Secondary education and higher	0.43	0.26	0.60
Urban	0.1	0.07	0.12
Rural	– 0.1	– 0.12	– 0.07
Male	– 0.04	– 0.05	– 0.02
Female	0.04	0.02	0.05

Table 2: Fixed effects for Zambia (effect coding)

Variable	Mean	10 % quantile	90 % quantile
Constant	0.1	0.04	0.16
Working	0.01	− 0.01	0.02
Not working	− 0.01	− 0.02	0.01
No education and incomplete primary education	− 0.17	− 0.21	− 0.14
Complete primary education and incomplete secondary education	− 0.06	− 0.09	− 0.03
Secondary education and higher	0.24	0.18	0.30
Urban	0.09	0.06	0.12
Rural	− 0.09	− 0.12	− 0.06
Male	− 0.06	− 0.07	− 0.04
Female	0.06	0.04	0.07

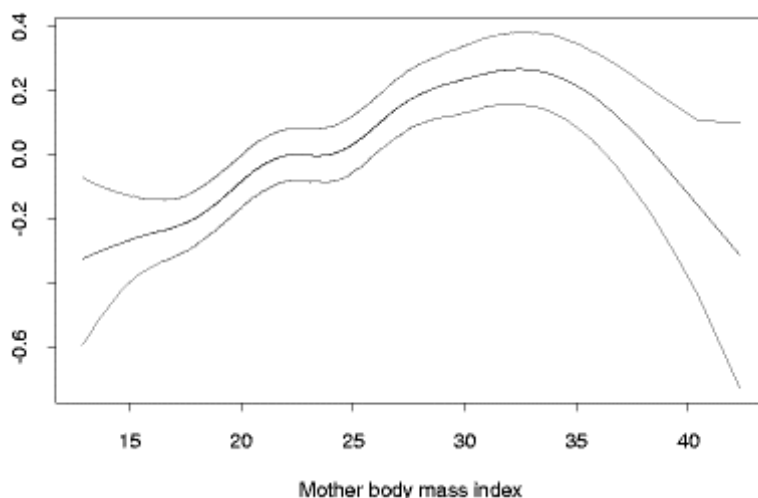


Figure 1: Non-linear effect of the mother's body mass index for Tanzania

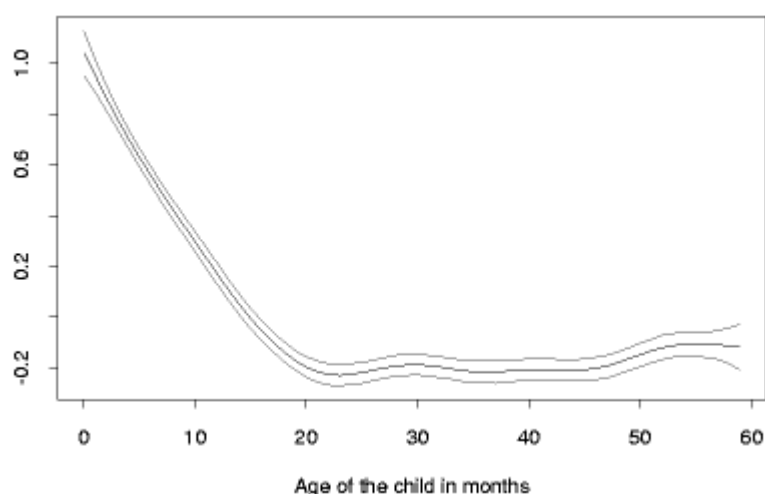


Figure 2: Non-linear effect of child's age for Tanzania

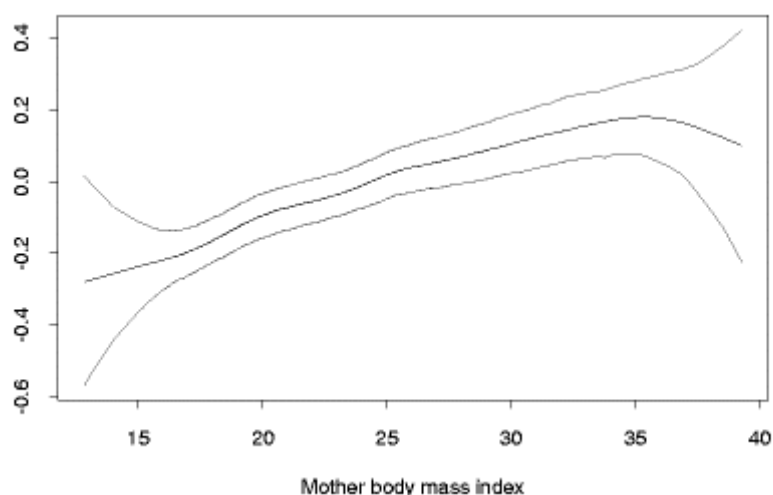


Figure 3: Non-linear effect of the mother's body mass index for Zambia

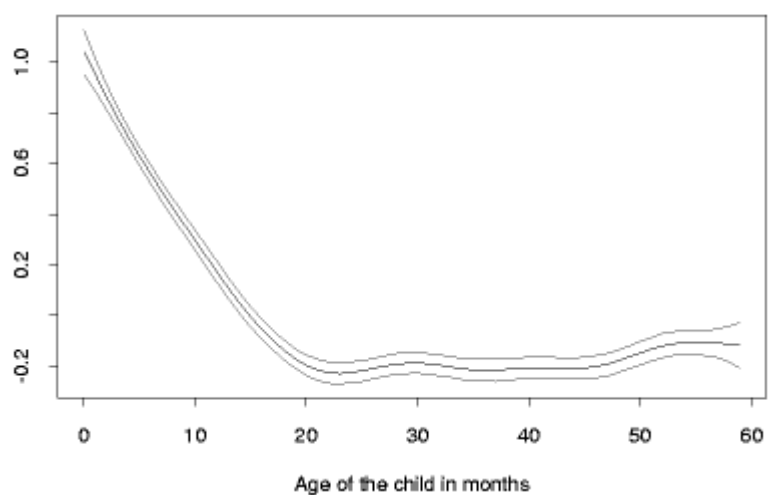


Figure 4: Non-linear effect of the child’s age for Zambia

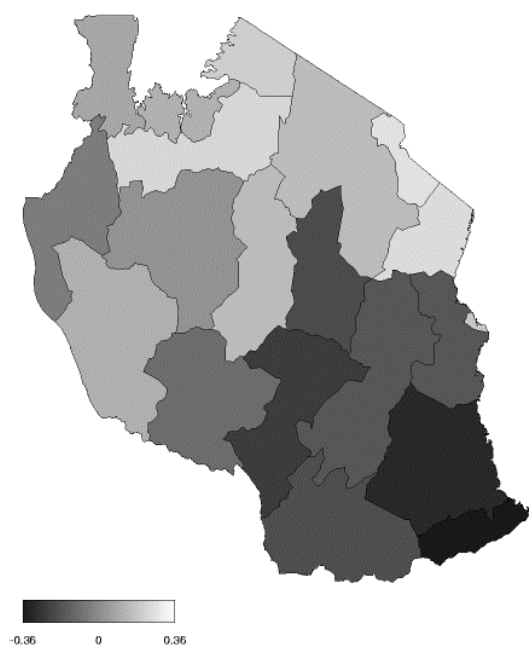


Figure 5: Posterior mean of the structured spatial effect for Tanzania

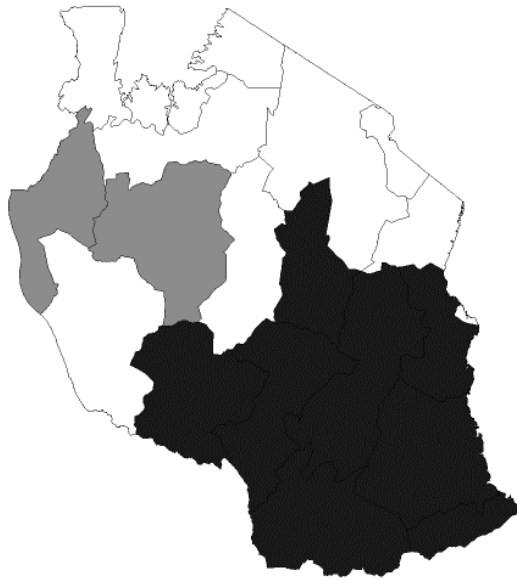


Figure 6: Posterior probabilities of the structured spatial effect for Tanzania

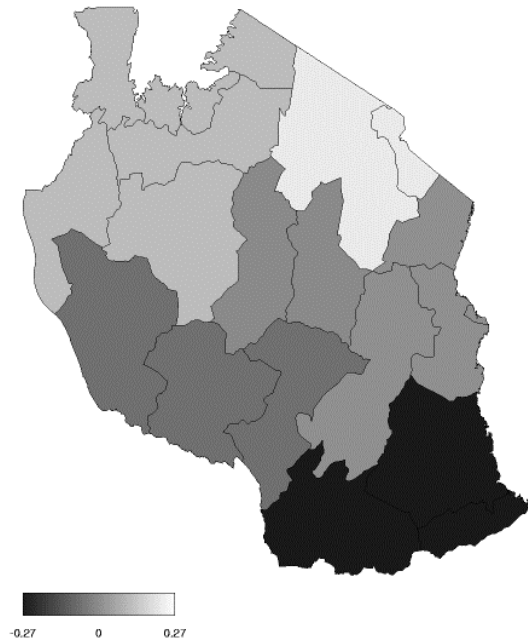


Figure 7: Regional effects for Tanzania

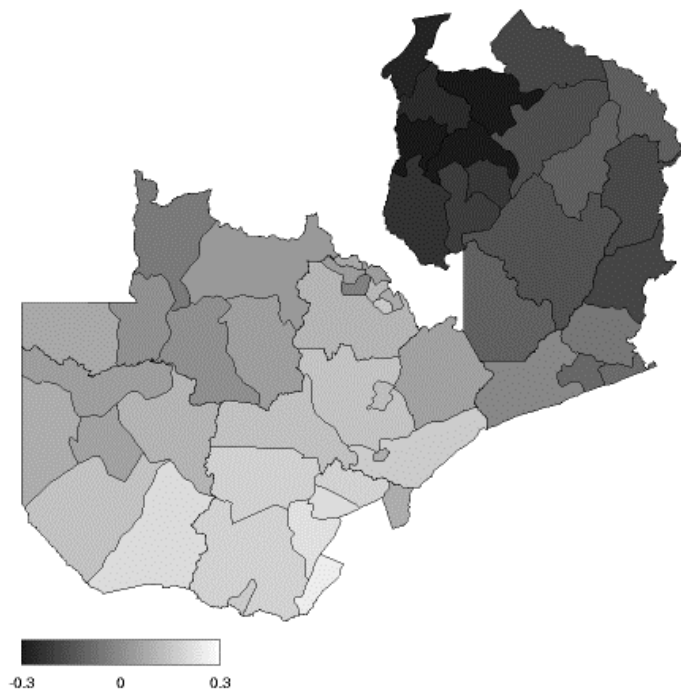


Figure 8: Posterior mean of the structured spatial effect for Zambia

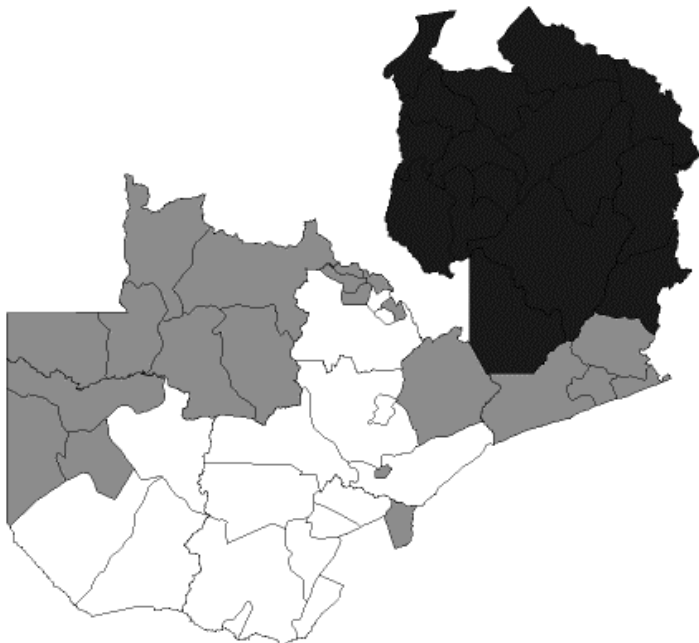


Figure 9: Posterior probabilities of the structured spatial effect for Zambia

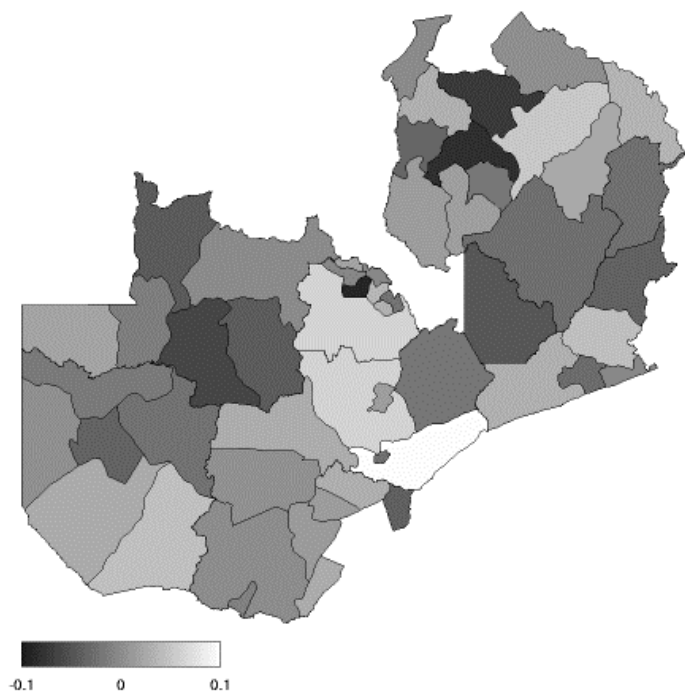


Figure 10: Posterior mean of the unstructured spatial effect for Zambia

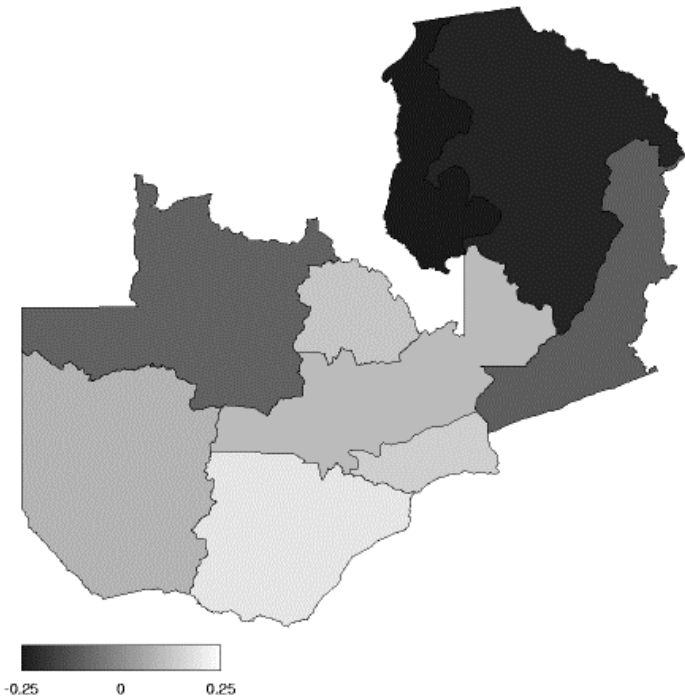


Figure 11: Regional effects for Zambia

7. References

- [1] Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995), ‘Bayesian computation and stochastic systems (with discussion)’, *Statistical Science*, Vol. 10, pp. 3–66.
- [2] Besag, J., York, Y. and Mollie, A. (1991), ‘Bayesian image restoration with two applications in spatial statistics (with discussion)’, *Annals of the Institute of Statistical Mathematics*, Vol. 43, pp. 1–59.
- [3] Biller, C. (2000), Adaptive Bayesian regression splines in semi-parametric generalised linear models, *Journal of Composition and. Statistical and Graphraphic Statistics*, Vol. 9, pp. 122–140.
- [4] De Boor, C. (1978), ‘A practical guide to splines’, Springer-Verlag, New York.
- [5] Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998), ‘Automatic Bayesian curve fitting’, *Journal of the Royal Statistical Society, B*, Vol. 60, pp. 333–350.
- [6] Eilers, P. and Marx, B. (1996), ‘Flexible smoothing with P-splines and penalised likelihood (with comments and rejoinder)’, *Statistical Science*, Vol. 11, pp. 89–121.
- [7] Fahrmeir, L. and Lang, S. (2001a), ‘Bayesian inference for generalised additive mixed models based on Markov random field priors, *Applied Statistics*, Vol. 50, pp. 201–220.
- [8] Fahrmeir, L. and Lang, S. (2001b), ‘Bayesian semi-parametric regression analysis of multicategorical time-space data’, *Annals of the Institute of Statistical Mathematics*, Vol. 53, pp. 11–30.
- [9] Gallup, J. and Sachs, J. (1998), ‘Geography and economic growth’, in Stiglitz, J. (ed.), *Proceedings of the Annual Bank Conference on Development Economics*, The World Bank, Washington, DC.
- [10] Hastie, T. and Tibshirani, R. (1993), ‘Varying-coefficient models’, *Journal of the Royal Statistical Society, Series B*, Vol. 55, pp. 757–796.
- [11] Hastie, T. and Tibshirani, R. (2000), ‘Bayesian backfitting’, *Statistical Science*, Vol. 15(3), pp. 196–213.
- [12] Kandala, N. B., Lang, S. and Klasen, S. (2001), ‘Semi-parametric analysis of childhood undernutrition in developing countries’, to appear in ISBA 2000 proceedings.
- [13] Klasen, S. (1996), ‘Nutrition, health, and mortality in sub-Saharan Africa: Is there a gender bias?’, *Journal of Development Studies*, Vol. 32, pp. 913–932.

- [14] Klasen, S. (1999), 'Malnourished and low mortality in South Asia, better nourished and dying young in Africa: What can explain this puzzle?', SFB 386 Discussion Paper No 214, University of Munich.
- [15] Lang, S. and Brezger, A. (2000), 'BayesX — Software for Bayesian inference based on Markov chain Monte Carlo simulation techniques' (available at <http://www.stat.uni-muenchen.de/~lang/>).
- [16] Lang, S. and Brezger, A. (2001), 'Bayesian P-splines', SFB 386 Discussion Paper No 236, University of Munich.
- [17] Moradi, A. and Klasen, S. (2000), 'The nutritional status of elites in India, Kenya, and Zambia: An appropriate guide for developing international reference standards for undernutrition?', SFB 386 Discussion Paper No 217, University of Munich.
- [18] Nyovani, J. M., Matthews, Z. and Margetts, B. (1999), 'Heterogeneity of child nutritional status between households: A comparison of six sub-Saharan African countries', *Population Studies*, Vol. 53, pp. 331–343.
- [19] Rue, H. (2001), 'Fast sampling of Gaussian Markov random fields with applications', to appear in *Journal of the Royal Statistical Society*, B.
- [20] Sen, A. (1999), *Development as freedom*, Oxford University Press, Oxford.
- [21] Smith, M. and Kohn, R. (1996), 'Non-parametric regression using Bayesian variable selection', *Journal of Econometrics*, Vol. 75, pp. 317–343.
- [22] Sommerfelt, A. E. and Stewart, M. K. (1994), 'Children's nutritional status', *Demographic and Health Surveys Comparative Studies*, No 12, Macro International.
- [23] Stephenson, C. (1999), 'Burden of infection on growth failure', *Journal of Nutrition, Supplement*, 534S–538S.
- [24] Svedberg, P. (1996), 'Gender bias in sub-Saharan Africa: Reply and further evidence', *Journal of Development Studies*, Vol. 32, pp. 933–943.
- [25] Unicef (1998), *The state of the world's children 1998: Focus on nutrition*, Unicef, New York.
- [26] WHO (1983), *Measuring change in nutritional status*, WHO, Geneva.
- [27] WHO (1995), 'Physical status: The use and interpretation of anthropometry', *WHO Technical Report Series*, No 854, WHO, Geneva.

- [28] WHO (1999), *Infant and young child growth: The WHO multicentre growth reference study*, Executive Board: Implementation of Resolutions and Decisions EB105/Inf.doc/1, WHO, Geneva.
- [29] World Bank (1995), *Zambia poverty assessment*, The World Bank, Washington, DC.

Analysis of aggregated data in survey sampling with application to fertiliser/pesticide usage survey

Jaeyong Lee (*)⁽¹⁾, Christopher Holloman (**), Alan F. Karr (***)
and Ashish P. Sanil (***)

(*) *Pennsylvania State University, University Park, PA, United States*

(**) *Duke University, Durham, NC, United States*

(***) *National Institute of Statistical Sciences, Research Triangle Park, NC, United States*

Keywords: survey sampling, aggregated data, confidentiality, Markov chain Monte Carlo

Abstract

In many cases, the public release of survey or census data at fine geographical resolution (for example, counties) may endanger the confidentiality of respondents. A strategy for such cases is to aggregate neighbouring regions into larger units that satisfy confidentiality requirements. An aggregation procedure employed in a prototype system for the US National Agricultural Statistics Service is used as a context to investigate the impact of aggregation on statistical properties of the data. We propose a Bayesian simulation approach for the analysis of such aggregated data. As a consequence, we are able to specify the type of additional information (such as certain sample sizes) that needs to be released in order to enable the user to perform meaningful analyses with the aggregated data.

1. Introduction

The work presented here derives from issues encountered at the National Institute of Statistical Sciences (NISS) in the course of developing a web-based system for disseminating survey data collected by the US National Agricultural Statistics Service (NASS).

This system was designed for the case wherein the public release of data (on use of agricultural chemicals) at the county level may compromise the confidentiality of survey respondents. The mechanism adopted for preserving confidentiality is geographical aggregation: data from adjacent non-disclosable counties are aggregated to the level of disclosable ‘supercounties.’ See Karr et al. (2001) for details.

While aggregation prevents disclosure, it may also, however, distort the data. Thus, the central question we address is: To what extent does confidentiality-preserving aggregation hamper a statistician’s ability to make informative inferences about the surveyed population?

⁽¹⁾ Research supported by NSF grants EIA-9876619 and DMS-9711365 to the National Institute of Statistical Sciences.

The main contribution is a Bayesian simulation approach for the analysis of such aggregated data. As a consequence, we are able to specify the type of additional information (such as certain sample sizes) that the disseminator needs to release in order to enable the user to perform meaningful analyses with the aggregated data.

We begin, in Section 2, with a description of the NASS data and disclosure problem. We outline the disclosure-risk criteria adopted by NASS (the so-called (n,p) -rule — see Willenborg and de Waal, 1996), and sketch the aggregation procedure used in the NASS system. In Section 3, we present design-based estimators for the population mean and its variance estimator, and show that they are unbiased. Within the setting of the design-based approach, it is difficult to account for the effect of the (n,p) -rule as a risk criterion. As an alternative, we propose in Section 4 a Bayesian approach to the problem; a companion simulation study is presented in Section 5. In Section 6, we present a Bayesian analysis in the NASS survey setting. Conclusions appear in Section 7.

2. The NASS system

2.1. Data

The data, which are collected by NASS through an annual survey of farms, consist of on-farm use of various chemicals (fertilisers, insecticides, herbicides and fungicides) on various crops. For our purposes, each data record can be thought of as consisting of farm identification (ID), farm size in acres, crop, chemical, pounds of chemical applied, state, county and year.

The primary information reported by NASS is application rates (pounds applied per acre) of certain chemicals on particular crops in geographical regions of interest. Ideally, for instance, a user would be able to learn the rate of application of the herbicide Alachlor on cornfields in all counties in North Carolina in 1996.

2.2. Disclosure criteria

NASS is concerned about protecting the identities (specifically the sampling weights) of farms in the survey. Such a disclosure would breach the respondent confidentiality promised by NASS, and would be even more alarming if some action were taken against the farm (even erroneously), such as litigation for excessive soil contamination or harm to workers. Thus, any information that would enable a user to estimate accurately the chemical usage on a particular farm is considered undisclosable (Dalenius, 1977).

NASS employs two rules to determine if the data for a particular county pose a confidentiality risk. The first is the n -rule: if only one or two farms were sampled in the county (thus, $n = 3$), the possibility of re-identification is too high, and the application rate for that county is undisclosable.

Second is the p -rule: a county-level application rate is undisclosable if the sample contains a dominant farm, the size of which (in acres) is more than p % of the total size of all farms sampled in that county (the system built by NISS uses $p = 60$ %). The rationale is that a

farm which dominates in a sample is susceptible to both identity and attribute disclosure risks.

We refer to these risk criteria collectively as the (n,p) -rule, which is a version of what is referred to in the statistical data disclosure literature as the (n,k) -rule (see Willenborg and de Waal, 1996). Note that there is an additional level of protection arising from the user's not knowing which farms are included in the survey.

2.3. The NASS system: aggregation to prevent disclosure

Currently, NASS releases chemical usage information only at the state level. At the other extreme, a system that simply refused queries for undisclosable counties would lead to unacceptable user annoyance and disillusionment, since, for the NASS data and the (n,p) -rule with $n = 3$ and $p = 60\%$, more than a half of the counties are undisclosable.

The system implemented by NISS produces intermediate aggregations that are more informative than state-level data but preserve confidentiality. It does so by aggregating undisclosable counties with neighbouring counties to form disclosable 'supercounties'. As a result, NASS can release data at the finest level of detail consistent with the risk criteria. Aggregations must be computed automatically, since there too many (crop, chemical, year, state) combinations to permit manual aggregation on a case-by-case basis. The system employs heuristic, 'greedy' algorithms based on the following procedure: examine the undisclosable (super)counties in a random order and merge them with a neighbouring (super)county according to some criterion for desirability of merging with the selected neighbour; continue until only disclosable (super)counties remain. See Karr et al. (2001), as well as the more detailed description in Karr et al. (2000), for specifics. A prototype version of the system is available on the web at <http://niss.cnidr.org>.

3. Design-based estimators

We abstract the NASS system as follows. Let $\mathcal{P} = (Y_1, Y_2, \dots, Y_N)$ be the population (of farms) of interest, where N is the size of the population. (Initially, we suppose that farms have only one attribute.) Suppose that the disseminator samples $\mathcal{S} = (y_1, y_2, \dots, y_n)$ using simple random sampling from \mathcal{P} .

The disseminator also draws a partition $\mathbf{\kappa} = (\kappa_1, \kappa_2, \dots, \kappa_k)$ of the index set $(1, 2, \dots, n)$ from a distribution $p(\mathbf{\kappa} | \mathcal{S})$. Often, $p(\mathbf{\kappa} | \mathcal{S})$ is independent of \mathcal{S} : for example, if κ_i is defined by geographical units such as counties or states, $\mathbf{\kappa}$ is independent of \mathcal{S} . On the other hand, in the NASS setting, $\mathbf{\kappa}$ depends on the sampled values y_i , which leads to the need for the Bayesian approach described in Section 4.

Let n_i be the number of observations in the i^{th} partition, κ_i for $i = 1, 2, \dots, k$. The disseminator aggregates the sample over the partition $\mathbf{\kappa}$ and releases

$$\mathcal{A} = \left\{ (\bar{y}_1, n_1), (\bar{y}_2, n_2), \dots, (\bar{y}_k, n_k) \right\},$$

where $\bar{y}_i = \sum_{j \in \kappa_i} y_j / n_i$.

3.1. Estimation of population mean

The sample mean $\bar{y} = \sum_{i=1}^n y_i / n$ is still available from the aggregated data \mathcal{A} and is a design-unbiased estimator regardless of whether $\mathbf{\kappa}$ is independent of \mathcal{S} .

The usual variance estimator $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$ of $S^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (N-1)$ cannot be recovered from \mathcal{A} in general. However, if $\mathbf{\kappa}$ is independent of the sample \mathcal{S} , a design-based unbiased estimator is available for S^2 .

Theorem 1: (a) The sample mean \bar{y} is a design-based unbiased estimator of the population mean for any distribution of $\mathbf{\kappa}$.
(b) If $\mathbf{\kappa}$ is independent of \mathcal{S} , then

$$s_a^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

is an unbiased estimator of $\text{Var}(\bar{y})$.

Proof. Since \bar{y} is independent of $\mathbf{\kappa}$, it is unbiased. Now, we prove s_a^2 is unbiased. First, note that

$$\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k n_i \left((\bar{y}_i - \bar{Y}) - (\bar{y} - \bar{Y}) \right)^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{Y})^2 - n (\bar{y} - \bar{Y})^2.$$

By Theorem 2.2 of Cochran (1977),

$$E(\bar{y} - \bar{Y})^2 = \left(1 - \frac{n}{N} \right) \frac{S^2}{n}.$$

Since $\mathbf{\kappa}$ is independent of \mathcal{S} , the conditional distribution of y_i given $\mathbf{\kappa}$ is the same as that of the sample mean of a simple random sample of size n_i , hence,

$$E\left[(\bar{y}_i - \bar{Y})^2 \mid \mathbf{\kappa} \right] = \left(\frac{N - n_i}{N} \right) \frac{S^2}{n_i}.$$

Combining these results, we get

$$E\left[\sum_{i=1}^k n_i (\bar{y}_i - \bar{Y})^2 \mid \mathbf{\kappa} \right] = \frac{S^2}{N} (kN - n) - S^2 \frac{N - n}{N} = (k-1)S^2.$$

This completes the proof.

Corollary 1: If κ is independent of \mathcal{S} , then

$$v_a(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s_a^2}{n}$$

is an unbiased estimator of $\text{Var}(\bar{y})$.

Note that the variance estimator in corollary 1 is not a valid estimator for aggregated data arising from the (n,p) -rule, because in this case the partition generated by it is not independent of \mathcal{S} . For this reason, we later introduce a Bayesian estimation procedure in which the (n,p) -rule can be accommodated in the analysis.

3.2. Estimation of ratios

Moving closer to the NASS setting, suppose that data now have two attributes: the population of interest is $\mathcal{P} = ((X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N))$. For the NASS problem, the X_i are farm sizes and the Y_i are amounts of chemical used. The application rates Y_i / X_i are the information to be released.

As in Section 3.1, let $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be a simple random sample from \mathcal{P} , and let κ be the partition of $\{1, 2, 3, \dots, n\}$ that defines the aggregated data. The disseminator then releases

$$\mathcal{A} = \{(\bar{x}_1, \bar{y}_1, n_1), (\bar{x}_2, \bar{y}_2, n_2), \dots, (\bar{x}_k, \bar{y}_k, n_k)\},$$

where $\bar{x}_i = \sum_{j \in \kappa_i} x_j / n_i$ and $\bar{y}_i = \sum_{j \in \kappa_i} y_j / n_i$.

The customary estimator for the population ratio $R = \bar{Y} / \bar{X}$ is

$$\hat{R} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

From the aggregated sample \mathcal{A} , the ratio estimator \hat{R} can be recovered, but (as in the case of population mean) its variance estimator based on \mathcal{S} cannot be recovered from \mathcal{A} .

From theorem 2.5 of Cochran (1977), the variance of the ratio estimator is approximately

$$\frac{(1 - n/N) \sum_{i=1}^N (Y_i - RX_i)^2}{nX^2(N-1)}.$$

Note that

$$\frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1}$$

is the population variance of quantities $D_i = Y_i - RX_i$, and can be estimated based on \mathcal{A} by

$$\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - R\bar{x}_i)^2.$$

This leads us to

$$v_a(\hat{R}) = \frac{(1-f) \sum_{i=1}^k n_i (\bar{y}_i - \hat{R}\bar{x}_i)^2}{\frac{-2}{n\bar{x}} (k-1)}$$

as an estimator for the variance of \hat{R} . As in the estimation of the population mean, this estimator is based on the assumption that κ is independent of \mathcal{S} . Note also that for the variance estimation of \hat{R} the sizes of aggregates (n_1, n_2, \dots, n_k) must be released.

4. Bayesian analysis of univariate data

In this section, we discuss the Bayesian analysis of aggregated data with the (n, p) -rule. Recall that the design-based estimators in Section 3 do not account for the effect of aggregations that are dependent on the sample.

Suppose the population (Y_1, Y_2, \dots, Y_N) is drawn from a parametric density $f(y|\theta)$, and (y_1, y_2, \dots, y_n) are a simple random sample from the population. Denote the sum of the y -values in the i th partition by y_i , i.e. $y_i = \sum_{j \in \kappa_i} y_j$. Then, based on aggregated data $(n_1, y_1), (n_2, y_2), \dots, (n_k, y_k)$, and the prior $\pi(\theta)$, the likelihood is

$$\pi(\theta) \prod_{i=1}^k f^{*n_i}(y_i | \theta) p(\kappa | \mathcal{S}),$$

where f^{*k} is the density of k -convolution of f . If f is a normal or gamma density, f^{*k} is known, but in many cases, of course, f^{*k} is unknown. This does not pose much difficulty in Bayesian computation with latent variables (y_1, y_2, \dots, y_N) in Markov chain Monte Carlo (MCMC) analyses.

The joint posterior distribution of $(\theta, \mathcal{P} \setminus \mathcal{S}, \mathcal{S})$ given \mathcal{A} is proportional to

$$\pi(\theta) p(\kappa | \mathcal{S}) \prod_{i=1}^N f(Y_i | \theta) \prod_{j=1}^k I\left(\sum_{l \in \kappa_j} y_l = y_j\right).$$

First, note that the distribution of the unsampled portion of the population, $(Y_{n+1}, Y_{n+2}, \dots, Y_N)$ can be integrated out from the posterior distribution and it suffices to

consider only the likelihood of the sampled data. Moreover, if κ is independent of \mathcal{S} , the factor $p(\kappa)$ does not contribute to the likelihood, and can be dropped. On the other hand, if κ is dependent on \mathcal{S} — which happens for aggregations produced with the (n,p) -rule — $p(\kappa | \mathcal{S})$ needs to be considered in the likelihood calculation.

In this paper, we focus on aggregated data arising from the (n,p) -rule. In this case, and in many others, because of the complicated nature of aggregation algorithms, the probability distribution of κ , $p(\kappa | \mathcal{S})$ is unknown not only to the user but also to the disseminator.

In order to develop insight into the issues, in this particular MCMC computation, consider a simpler case: we assume that the partition is drawn from a linear aggregation algorithm. It forms an aggregate by adding observations one by one until the aggregate satisfies the (n,p) -rule.

Linear aggregation algorithm

Step 1: Set $i \leftarrow 1$, and $\mathcal{A} \leftarrow \phi$.

Step 2: Repeat until \mathcal{A} satisfies the (n,p) -rule: $\mathcal{A} \leftarrow \mathcal{A} \cup \{\epsilon\}$; $i \leftarrow i + 1$.

Step 3: Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{A\}$.

Step 4: Set $\mathcal{A} \leftarrow \phi$. Go to step 2.

In practice, the last aggregate may not satisfy the (n,p) -rule. In this event, it is merged with the previous aggregate, and we check whether the (n,p) -rule is satisfied. If not, the process of backward aggregation is continued until it is. Because the aggregate constitutes only a very small portion of the whole data, this additional step will not be considered in the likelihood.

If κ is drawn to satisfy the (n,p) -rule with n_0 and p_0 , then the posterior distribution of (θ, \mathcal{P}) given \mathcal{A} is proportional to

$$\pi(\theta) \prod_{j=1}^n f(y_j | \theta) \prod_{i=1}^k \left[\prod_{l=m_{i-1}+1}^{m_i-1} I \left(\max_{m_{i-1}+1 \leq b \leq l} y_b > p_0 \sum_{b=m_{i-1}+1}^l y_b \right) \right] I \left(\max_{m_{i-1}+1 \leq b \leq m_i} y_b \leq p_0 \sum_{b=m_{i-1}+1}^{m_i} y_b \right)$$

where $m_0, m_i = \sum_{j=1}^i n_j$ for $i = 1, 2, \dots, k$.

The MCMC steps follow:

Step 1: Generation of θ given y_1, y_2, \dots, y_n .

The conditional distribution of θ is proportional to $\pi(\theta) \prod_{i=1}^n f(y_i | \theta)$. Hence, sampling θ is the same as sampling θ from the posterior of i.i.d. observations, y_1, y_2, \dots, y_n .

Step 2: Generation of y_1, y_2, \dots, y_n , given \mathcal{A} and θ .

We illustrate this step with the first aggregate, for the rest is the same. First, generate y_1, y_2, \dots, y_{n_0} by the acceptance–rejection sampling methods, so that $y_i > p_0 \sum_{j=1}^{n_0} y_j$ for some $i = 1, 2, \dots, n_0$. Second, for $l = n_0 + 1, \dots, n_1 - 1$, generate y_l by the acceptance–rejection or inverse-cdf sampling method, so that $y_i > p_0 \sum_{j=1}^l y_j$ for some $i = 1, 2, \dots, l$. Third, sample y_{n_1} by the acceptance–rejection or inverse-cdf (cumulative distribution function) sampling method so that $y_i \leq p_0 \sum_{j=1}^{n_1} y_j$ for all $i = 1, 2, \dots, n_1$.

If inference on the finite population y_1, y_2, \dots, y_N is necessary, the following additional step can be added to sample the unsampled portion of the finite population.

Step 3: Sampling y_{n+1}, y_2, \dots, y_N , given θ .

Sample y_{n+1}, y_2, \dots, y_N that are i.i.d. from $f(y | \theta)$.

Finally, note here that we need to know the value of p in the (n, p) -rule in order to construct the MCMC algorithm and the likelihood on which the posterior is based. Otherwise, the Bayesian analysis is simply not possible. For this reason, in disseminating aggregated data, the numbers of subjects in each aggregate as well as the value of p must be released to users.

5. Simulation study

While the design-based approach gives simple mean and ratio estimators, it cannot accommodate the variability arising from partitions dependent on the data, as occurs for the (n, p) -rule. Since the dependence is caused only by the ‘ p portion’ of the (n, p) -rule (an n -rule alone leads to aggregations dependent on the sampling, but not the sampled data values), the simulation study focuses on the effect of the p -rule on confidence intervals for design-based and Bayesian approaches.

The 500 samples of size 100 are generated from Gamma(1,1) and Gamma(0.5,0.5), with population means 1,1 and variances 1,2, respectively. Since the variance affects the size of aggregation with the p -rule, n is fixed at 3 in all simulations. For each sample of size 100, aggregates were generated by the linear aggregation algorithm, and, for each aggregated sample, 95 % design-based and Bayesian credible intervals for the population mean and their lengths are computed, and then checked if they contain the population mean. The design-based confidence interval is computed with

$$\bar{x} \pm 1.96 \nu_a(\bar{x})$$

The exponential distribution with mean 100 is employed for the gamma parameters α and β , reflecting vague prior information. We tried exponential priors with different means, but

the posterior seems to be robust with respect to the choice of the prior parameters. The number of MCMC iterations used is 5 000, which seems enough for the model.

The simulation results are shown in Table 1. Since the design-based confidence interval does not account for the p -rule, the coverage probability is slightly less than the nominal coverage rate of 95 % and the lengths remain about the same as p varies. This is expected because the design-based approach does not take the p -rule into consideration and does not assume the parametric distributional form of the population.

In fact, we initially expected poorer behaviour of the design-based approach, but, to our surprise, its performance is not as bad as we expected.

The coverage probabilities of Bayesian credible intervals fluctuate around the nominal coverage probability and the lengths of the credible sets increase as p decreases. There is a tendency for the simulated coverage probability of the Bayesian credible set to increase as p decreases. We are not sure what causes this.

Table 1: Simulated coverage probabilities and lengths of the confidence intervals for \bar{X} for the n -rule with $n = 3$ — samples are drawn from Gamma (α, α)

Population	p -rule	Design-based confidence interval		Bayesian credible interval	
		Coverage probability	Length	Coverage probability	Length
Gamma(1,1)	0.9	0.908	0.386	0.934	0.397
	0.8	0.934	0.385	0.948	0.403
	0.7	0.946	0.384	0.962	0.406
	0.6	0.928	0.387	0.950	0.422
	0.5	0.920	0.380	0.966	0.452
	0.9	0.920	0.547	0.946	0.566
	0.8	0.934	0.550	0.948	0.570
	0.7	0.936	0.539	0.954	0.580
	0.6	0.912	0.534	0.958	0.601
	0.5	0.916	0.546	0.970	0.664

6. Analysis of NASS chemical usage survey

In this section, we analyse data in the NASS setting, which were aggregated using the algorithms presented in Karr et al. (2000).

6.1. The model, the prior and the posterior

For a given crop (vegetable or fruit) and chemical active ingredient (AI), the population is $\mathcal{P} = ((X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N))$.

where X_i is the size of the i th farm in acres and Y_i is the amount of AI used on that farm. We assume X_i are i.i.d. from $\text{Gamma}(\alpha, \beta)$. Since Y_i is the amount of AI applied to the land of size X_i , it is natural to assume that $Z_i = Y_i/X_i$ are exchangeable; hence, we assume that Z_i are i.i.d. from $\text{Gamma}(\gamma, \delta)$.

The simple random sample $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ is drawn from the population. The partition κ is assumed to be formed by the linear aggregation algorithm with the (n, p) -rule on x_i with $n = n_0$ and $p = p_0$. The data released, therefore, are

$$\mathcal{A} = \{(n_1, x_{1.}, y_{1.}), (n_2, x_{2.}, y_{2.}), \dots, (n_k, x_{k.}, y_{k.})\}.$$

We use very flat priors on the parameters: $\alpha \sim \text{Exponential}(\lambda_\alpha)$, $\beta \sim \text{Exponential}(\lambda_\beta)$, $\gamma \sim \text{Exponential}(\lambda_\gamma)$ and $\delta \sim \text{Exponential}(\lambda_\delta)$, where $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_\delta = 0.01$. The posterior distribution of $\alpha, \beta, \gamma, \delta$, and \mathcal{P} is proportional to

$$\begin{aligned} & \pi(\alpha)\pi(\beta)\pi(\gamma)\pi(\delta) \prod_{j=1}^n \text{Gamma}(x_j | \alpha, \beta) \prod_{j=1}^n \text{Gamma}(z_j | \gamma, \delta) \\ & \times \\ & \prod_{i=1}^k \left[\prod_{l=m_{i-1}+1}^{m_i-1} I\left(\max_{m_{i-1}+1 \leq b \leq l} x_b > p_0 \sum_{b=m_{i-1}+1}^l x_b\right) \right] I\left(\max_{m_{i-1}+1 \leq b \leq m_i} x_b \leq p_0 \sum_{b=m_{i-1}+1}^{m_i} x_b\right) \\ & \times \prod_{i=1}^k I\left(\sum_{j=m_{i-1}+1}^{m_i} x_j = x_i\right) I\left(\sum_{j=m_{i-1}+1}^{m_i} x_j z_j = y_i\right). \end{aligned}$$

Note that the unsampled portion of the population is integrated out from the posterior distribution.

The MCMC steps are as follows:

Sampling α , given $\beta, \gamma, \delta, \mathbf{X}$ and \mathbf{Z}

The full conditional distribution of α is proportional to

$$e^{-\lambda_\beta \beta} \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} \propto \frac{1}{\Gamma(\alpha)^n} \exp \left[-\alpha \left(\lambda_\alpha - n \log \beta - \sum_{i=1}^n \log x_i \right) \right]$$

This is not a well-known distribution, and we use a random walk Metropolis-Hastings sampler with a uniform proposal distribution.

Sampling γ , given $\alpha, \beta, \delta, \mathbf{X}$ and \mathbf{Z}

The full conditional of γ is of the same form as that of α . It is proportional to

$$\frac{1}{\Gamma(\gamma)^n} \exp \left[-\gamma \left(\lambda_\gamma - n \log \delta - \sum_{i=1}^n \log z_i \right) \right].$$

A random walk Metropolis-Hastings sampler with a uniform proposal distribution is employed.

Sampling β , given $\alpha, \gamma, \delta, \mathbf{X}$ and \mathbf{Z}

This full conditional turns out to be a gamma distribution:

$$\text{Gamma}\left(n\alpha + 1, \lambda_\beta + \sum_{i=1}^n x_i\right).$$

Sampling δ , given $\alpha, \beta, \gamma, \mathbf{X}$ and \mathbf{Z}

As in the full conditional distribution of β , the conditional distribution of δ is

$$\text{Gamma}\left(n\gamma + 1, \lambda_\delta + \sum_{i=1}^n z_i\right).$$

Sampling $x_j, z_j, j \in \kappa_i$, given $\alpha, \beta, \gamma, \delta, \sum_{j \in \kappa_i} x_j = x_i$ and $\sum_{j \in \kappa_i} x_j z_j = y_i$.

For simplicity of notation, we will consider only κ_1 and suppose

$\kappa_1 = \{1, 2, \dots, n_1\}$. Hence, we need to sample $x_1, x_2, \dots, x_{n_1}, z_1, z_2, \dots, z_{n_1}$, conditional on $x_1, y_1, \alpha, \beta, \gamma, \delta$. Their conditional distribution, however, is not of a simple form. We therefore sample x_1, x_2, \dots, x_{n_1} and $y_1 = x_1 z_1, y_2 = x_2 z_2, \dots, y_{n_1} = x_{n_1} z_{n_1}$. The latter conditional distribution is proportional to

$$\begin{aligned} & \prod_{i=1}^{n_1} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} \\ & \times \prod_{j=1}^{n_1} I\left(\max_{1 \leq b \leq n_1} x_b > p_0 \sum_{b=1}^j x_b\right) I\left(\max_{1 \leq b \leq n_1} x_b \leq p_0 \sum_{b=1}^{n_1} x_b\right) I\left(\sum_{i=1}^{n_1} x_i = x_1\right) \\ & \times \prod_{i=1}^{n_1} \frac{(\delta/x_i)^\gamma}{\Gamma(\gamma)} y_i^{\gamma-1} e^{-\delta y_i/x_i} I\left(\sum_{i=1}^{n_1} y_i = y_1\right). \end{aligned}$$

First, x_1, x_2, \dots, x_{n_1} are sampled in the same way as step 2 in Section 4. The difficult part is to sample $y_1 = x_1 z_1, y_2 = x_2 z_2, \dots, y_{n_1} = x_{n_1} z_{n_1}$ conditional on x_1, x_2, \dots, x_{n_1} and y_1 . We use a Metropolis-Hastings algorithm with the Dirichlet distribution with parameter $(\gamma^*, \gamma^*, \gamma^*, \dots, \gamma^*)$. The acceptance rule hence becomes

$$\begin{aligned}\alpha((X, Y), (X', Y')) &= \frac{\prod_{i=1}^{n_1} (y'_i/x'_i)^{\gamma-1} e^{-\delta \sum_{i=1}^{n_1} y'_i/x'_i} \prod_{i=1}^{n_1} y_i^{\gamma^*-1} e^{-\sum_{i=1}^{n_1} y_i}}{\prod_{i=1}^{n_1} (y_i/x_i)^{\gamma-1} e^{-\delta \sum_{i=1}^{n_1} y_i/x_i} \prod_{i=1}^{n_1} (y'_i)^{\gamma^*-1} e^{-\sum_{i=1}^{n_1} y'_i}} \\ &= \left(\frac{y'_i x_i}{y_i x'_i} \right)^{\gamma-1} e^{-\delta \sum_{i=1}^{n_1} (y'_i/x'_i - y_i/x_i)} \prod_{i=1}^{n_1} \left(\frac{y_i}{y'_i} \right)^{\gamma^*-1}.\end{aligned}$$

The following two steps can be added if the unsampled portion of the population is necessary for the computation of the parameter of interest.

Sampling $y_i, i = n+1, \dots, N$, given $\alpha, \beta, \gamma, \delta, x_1, x_2, \dots, x_n$ and \mathbf{Z}

For $i = n+1, \dots, N$, the full conditional distribution of y_i is $\text{Gamma}(\alpha, \beta)$.

Sampling $z_i, i = n+1, \dots, N$, given $\alpha, \beta, \gamma, \delta, z_1, z_2, \dots, z_n$ and \mathbf{Z}

For $i = n+1, \dots, N$, the full conditional distribution of y_i is $\text{Gamma}(\gamma, \delta)$.

6.2. Posterior analysis

The Markov chain was run for 30 000 iterations, which seems ample for the convergence. The first 6 000 posterior samples were discarded for burn-in and from the remaining 24 000 samples every third output was used in the analysis. Histograms are based on the posterior samples of population mean (α/β) in acreage, α and β , ratio (γ/δ) in pounds per acre, γ and δ . Figure 1 shows the histograms of the six quantities and the summary statistics from the posterior samples of the six quantities displayed in Table 2.

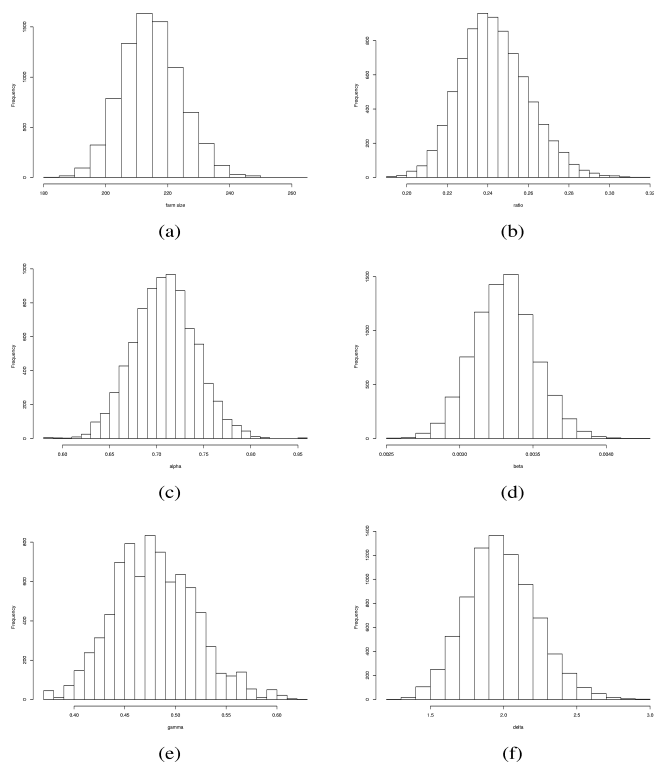


Figure 1: Histograms from the MCMC output: (a) population mean of farm sizes; (b) population mean of ratios; (c) α ; (d) β ; (e) γ ; (f) δ

Table 2: Summary statistics from the posterior distribution

Parameter	Mean	Standard deviation	95 % credible set
Mean of farm size	214.8	9.541	(196.84,234.10)
Mean of ratio	0.243	0.0169	(0.213,0.279)
α	0.712	0.0360	(0.647,0.786)
β	0.00332	0.000223	(0.00291,0.00379)
γ	0.494	0.0584	(0.406,0.635)
δ	2.056	0.322	(1.551,2.794)

7. Conclusions

In this paper, design-based and Bayesian analyses of aggregated data are discussed. Design-based estimators are simple and easy to calculate, but the sample-dependent aggregation is not easy to consider in the analysis. As an alternative to the design-based analysis, the Bayesian analysis is discussed. Its main advantage over the design-based approach is that it can accommodate the variation due to the sample-dependent aggregation.

It is also discussed that in disseminating aggregated data using the (n,p) -rule the disseminator needs to release the level of p employed in the aggregation and the size of aggregates; otherwise, reasonable statistical analysis of the data may not be possible.

8. References

- [1] Blien, U., Wirth, H. and Müller, M. (1992), 'Disclosure risk for microdata stemming from official statistics', *Statistica Neerlandica*, Vol. 46, pp. 69–82.
- [2] Cochran, W. G. (1977), *Sampling techniques*, Third edition, Wiley, New York.
- [3] Dalenius, T. (1977), 'Toward a methodology for statistical disclosure control', *Statistik Tidskrift*, Vol. 5, pp. 429–444.
- [4] Duncan, G. T. and Lambert, D. (1986), 'Disclosure-limited data dissemination (with discussion)', *Journal of the American Statistical Association*, Vol. 81, pp. 10–28.
- [5] Fienberg, S. E., Makov, U. E. and Sanil, A. P. (1997), 'A Bayesian approach to data disclosure: optimal intruder behaviour for continuous data', *Journal of Official Statistics*, Vol. 13, pp. 75–89.

- [6] Fuller, W. (1993), 'Masking procedures for microdata disclosure limitation', *Journal of Official Statistics*, Vol. 9, pp. 383–406.
- [7] Karr, A. F., Lee, J., Sanil, A., Hernandez, J., Karimi, S. and Litwin, K. (2000), 'Web-based systems that disseminate information from data but preserve confidentiality', Technical report, National Institute of Statistical Sciences (available online at www.niss.org/dg/technicalreports.html).
- [8] Karr, A. F., Lee, J., Sanil, A., Hernandez, J., Karmi, S. and Litwin, K. (2001), 'Disseminating information but protecting confidentiality', *IEEE Computer*, Vol. 34(2), pp. 36–37.
- [9] Paass, G. (1988), 'Disclosure risk and disclosure avoidance for microdata', *Journal of Business and Economic Statistics*, Vol. 6, pp. 487–500.
- [10] Willenborg, L. and de Waal, T. (1996), *Statistical disclosure control in practice*, Springer-Verlag, New York.

Bayesian multivariate micro-aggregation under the Hellinger's distance criterion

George Kokolakis (*)and Photis Nanopoulos (**)

(*)*Department of Mathematics, NTUA, Greece*

(**) *European Commission, Eurostat, Luxembourg*

Keywords: statistical confidentiality, statistical disclosure control, data masking, Hellinger's distance, identity protection, multivariate micro-aggregation, optimal partitioning

Abstract

The technique of micro-aggregation is widely used in Official Statistics to reduce the risk of identity disclosure. Micro-aggregation implies loss of information. In order to minimise the loss of information while protecting confidentiality of individual data, we present here a Bayesian method of optimal multivariate micro-aggregation by minimising the Hellinger's distance criterion between the two posteriors based on the original and the micro-aggregated multivariate normal data.

1. Introduction

One of the main problems that National Statistical Offices face when disseminating information to the public is that they are not allowed to deliver, even anonymous individual records, if there is any possibility of direct or indirect identification of individual units. This is particularly true for business data. Their task is to protect the individual data from what Duncan and Lambert (1989) have called 'the risk of identity disclosure'. This basic requirement of Official Statistics frequently has negative implications in pursuing statistical studies and thus researchers complain about the lack of microdata. Duncan and Pearson (1991) state that 'a number of social science research and public policy studies could be pursued if the present tension between access and confidentiality were better resolved'. For more on this topic see also Fienberg (1994).

In order to resolve the above problem several statistical techniques have been proposed for identity protection. These techniques can be classified, as in McGuckin and Nguyen (1988), into three main categories: (a) noise introduction, (b) data swapping and (c) micro-aggregation. In category (a), an error term with mean zero and known variance is added to each response variable in all records in the microdata file. In category (b), new synthetic records are constructed by switching blocks of information. In category (c), clusters of similar 'structure' are formed and then the averages of all records in each cluster form the micro-aggregate records.

Two additional techniques, not entirely belonging to one of the above categories but often used in practice, are the following: The 'data suppression' technique, where entire records are suppressed, and the 'recoding method', where new larger categories are formed by collapsing categories of one or more qualitative variables in the original microdata file.

It is useful to mention here a more recent method that unifies the above categories. This is the matrix-masking technique of Duncan and Pearson (1991). A triplet of matrices $(M) = (A, B, C)$, called the mask, is used to transform the original microdata file X into the masked microdata file $X_{(M)} = A X B + C$. The matrices A, B and C have elements that are either constants or random variables, and they are not necessarily independent of X . The problem of course is how to choose the mask (M) in order to preserve in $X_{(M)}$ the maximum possible information from X while protecting confidentiality.

Methods, such as those described above, quite often suffer from several defects. Firstly, they do not always provide satisfactory assurance that confidentiality will be preserved, especially with asymmetrical data distributions such as those found in the field of business statistics. Secondly, they introduce bias in the estimation of certain parameters (Adam and Wortmann, 1989). An additional defect is that the correlation structure of the original microdata can be seriously distorted.

The idea of micro-aggregation had been introduced in Eurostat by Photis Nanopoulos in 1991 as an attempt to solve the confidentiality problems in business panels. In a paper presented in the Statistics Canada Symposium by Defays and Nanopoulos (1992), it was proved that optimal k -size aggregates can be obtained by hyperplanes that might be chosen to be perpendicular to the lines joining the centres of gravity of each pair of k -size clusters. In a second paper, Defays *et al.* have generalised and improved the algorithm for these results. The criterion used for optimal partitioning in the above papers was the usual minimisation of the within-group Euclidean squared distance, a method often applied in the area of unsupervised learning and clustering, see Duda and Hart (1973), pp. 211–228.

The above technique is one of the matrix-masking type, with $B = I$, $C = 0$ and A depending on the data set X . But, since Euclidean distance takes into account neither different variabilities among the components of the records nor their covariance structure, it is a good partitioning criterion for isotropic data only.

What we propose here is to apply a Bayesian approach and minimise a measure of distance between two posterior distributions conditional on X and $X_{(M)}$ respectively.

2. Micro-aggregation with multivariate normal data

Since the rules do not permit the transmission of aggregated data when the number of aggregated records is less than a threshold value k_0 (in most countries $k_0 = 3$), we propose to apply this rule and replace individual data by averages of small aggregates. The problem is to partition the whole data set into clusters with sizes that satisfy the above constraint, while preserving the statistical information content of the original data set as

much as possible. The problem of assessing the associated risk of identity disclosure is not considered here as its study needs more contextual specifications. But, since increasing the cluster-size decreases the risk of identity disclosure, clusters of greater size have to be taken if we feel that some individuals are at a high risk of been identified. For this kind of problem, see Omori (1999), Samuels (1998) and Fienberg and Makov (1998).

Suppose that the microdata file consists of n individual records x_i ($i = 1, \dots, n$). Each x is a d -vector $(x_1, \dots, x_d)^T$ that has a multivariate Normal distribution with mean μ and precision matrix r . Both mean μ and r are considered unknown random quantities. The x 's, conditional on μ and r , are independent. The prior for μ and r is taken from the conjugate family, i.e. the multivariate Normal-Wishart family of distributions. Jeffreys' non-informative prior will be also considered as a limiting case of Normal-Wishart distributions. Suppose now the data set $X = \{x_i, i = 1, \dots, n\}$ is partitioned into groups G_l ($l = 1, \dots, L$) with k_l elements respectively. The mean of the group G_l forms the aggregate datum \bar{x}_l considered to have multiplicity k_l . The set so derived is denoted by $X_{(M)} = \{(\bar{x}_l, k_l), l = 1, \dots, L\}$.

The group means \bar{x}_l simply serve as approximations to the original x 's that belong to the same group G_l . Therefore we shall keep the independence property of the original x 's for all k_l records which have been set equal to \bar{x}_l . Consequently, when constructing the likelihood functions, both samples X and $X_{(M)}$ will have the same size n .

The problem is to divide the original data set X in such a way that the posterior distribution of $\theta = (\mu, r)$, based on the masked data set $X_{(M)}$, will be as 'close' as possible to their posterior distribution based on the original data set X .

3. Distances between distributions

There are several measures of proximity, distance or divergence between two probability distribution functions F_1 and F_2 in the literature. McLachlan (1992) p. 22, following Krzanowski (1983), classifies them into two major categories: (a) measures based on information theory ideas and (b) measures based on Bhattacharyya's measure of 'affinity' between two distributions. Assuming that the two distribution functions F_1 and F_2 admit densities f_1 and f_2 respectively with respect to some measure ν we have:

- (1) Kullback-Leibler measure of divergence

$$J(F_1, F_2) = \int f_1(\theta) \log \{f_1(\theta) / f_2(\theta)\} d\nu, \quad (1)$$

And

(2) Hellinger's distance

$$H(F_1, F_2) = \left[\int \left\{ \sqrt{f_1(\theta)} - \sqrt{f_2(\theta)} \right\}^2 d\nu \right]^{\frac{1}{2}} = (2 - 2\rho)^{\frac{1}{2}}, \quad (2)$$

where ρ is Bhattacharyya's measure of affinity defined by

$$\rho = \int \{f_1(\theta)f_2(\theta)\}^{\frac{1}{2}} d\nu. \quad (3)$$

In many cases J and H provide equivalent results. In this paper the optimality criterion we chose is minimising Hellinger's distance $H(F_1, F_2)$, or equivalently maximising Bhattacharyya's measure of affinity ρ .

The optimal partitioning of the original data set X will be derived by maximising the measure of affinity ρ between the two posterior distributions of $\theta = (\mu, r)$, the first conditional on the full data set X , and the other conditional on the masked data set $X_{(M)} = \{(\bar{x}_l, k_l), l = 1, \dots, L\}$ corresponding to a partition of X . Thus, we have to maximise with respect to the partition $G = \{G_1, \dots, G_L\}$ the quantity

$$\rho = \rho(G) = \int \{p(\mu, r | X)p(\mu, r | X_{(M)})\}^{1/2} d\mu dr. \quad (4)$$

Distributional assumptions

Given the mean μ and the precision matrix $r (= \Sigma^{-1})$, the x 's are independent, $N_d(\mu, r)$ distributed random d -vectors.

The joint prior for μ and r is as follows:

(a) Using a Normal-Wishart prior

Given the precision matrix r , the mean μ is $N_d(\mu_0, n_0 r)$ with $n_0 > 0$, and the precision matrix r has a Wishart distribution with degrees of freedom α_0 ($\alpha_0 > d-1$) and 'precision' matrix τ_0 . Specifically, cf. De Groot, (1970), pp. 177-179, or Robert (1994), pp. 151-155, we have:

$$p(r) = c \left| \tau_0 \right|^{\frac{\alpha_0}{2}} \left| r \right|^{\frac{\alpha_0 - d - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\tau_0 r) \right\}, \quad (5)$$

where

$$c = \left[2^{\frac{\alpha_0 d}{2}} \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^d \Gamma\left(\frac{\alpha_0 + 1 - j}{2}\right) \right]^{-1}.$$

Now, the two posterior distributions of (μ, r) , given the data set X or the data set $X_{(m)}$ respectively, are again multivariate Normal-Wishart with parameters:

$$\mu_j = \frac{n_0 \mu_0 + n m_j}{n_j}, \quad n_j = n_0 + n, \quad \alpha_j = \alpha_0 + n,$$

and

$$\tau_j = \tau_0 + S_j + \frac{n_0 n}{n_0 + n} (m_j - \mu_0)(m_j - \mu_0)^T, \quad j=1, 2, \quad (6)$$

where

$$m_1 = \bar{x}, \quad \text{and} \quad m_2 = \bar{x}_{[m]} = \bar{x}, \quad (7)$$

$$S_1 = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T, \quad (8)$$

$$S_2 = \sum_{l=1}^L k_l (\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})^T, \quad (9)$$

Note that S_1 and S_2 are like the total and between-group scatter matrices in a one-way MANOVA setting. Since from the above we have $m_1 = m_2$, we also have:

$$\mu_1 = \mu_2 \equiv \mu^*, \quad n_1 = n_2 \equiv n^*, \quad \alpha_1 = \alpha_2 \equiv \alpha^*.$$

In addition

$$\tau_1 - \tau_2 = S_1 - S_2 \equiv S_3, \quad (10)$$

where

$$S_3 = \sum_{l=1}^L \sum_{x_i \in G_l} (x_i - \bar{x}_l)(x_i - \bar{x}_l)^T, \quad (11)$$

the within-group scatter matrix.

Introducing the above results into (4) we get:

$$\begin{aligned} \rho &= \iint \{N_d(\mu | \mu_1, n_1 r) N_d(\mu | \mu_2, n_2 r) W_d(r | \alpha_1, \tau_1) W_d(r | \alpha_2, \tau_2)\}^{\frac{1}{2}} d\mu dr \\ &= \iint N_d(\mu | \mu^*, n^* r) \{W_d(r | \alpha^*, \tau_1) W_d(r | \alpha^*, \tau_2)\}^{\frac{1}{2}} d\mu dr \\ &= \int \{W_d(r | \alpha^*, \tau_1) W_d(r | \alpha^*, \tau_2)\}^{\frac{1}{2}} dr \end{aligned}$$

and finally,

$$\rho = \frac{\{|\tau_1| |\tau_2|\}^{\frac{\alpha^*}{4}}}{|\frac{1}{2}(\tau_1 + \tau_2)|^{\frac{\alpha^*}{2}}}, \quad (12)$$

where $\alpha^* = \alpha_0 + n$. Since α^* does not depend on the partition, we can take the α^* -th root of the above result and get the following expression for the modified measure of affinity ρ^* :

$$\rho^* = \rho^{\frac{1}{\alpha^*}} = \frac{\{|\tau_1| |\tau_2|\}^{\frac{1}{4}}}{|\frac{1}{2}(\tau_1 + \tau_2)|^{\frac{1}{2}}} = 2^{\frac{d}{2}} \frac{|\tau_1^{-1} \tau_2|^{\frac{1}{4}}}{|I + \tau_1^{-1} \tau_2|^{\frac{1}{2}}} = 2^{\frac{d}{2}} \frac{|\tau_1^{-\frac{1}{2}} \tau_2 \tau_1^{-\frac{1}{2}}|^{\frac{1}{4}}}{|I + \tau_1^{-\frac{1}{2}} \tau_2 \tau_1^{-\frac{1}{2}}|^{\frac{1}{2}}} \quad (13)$$

with the last equality based on the fact that τ_1 and τ_2 are positive definite symmetric matrices.

If λ_j ($j = 1, \dots, d$) are the eigenvalues of the positive definite matrix $\tau_1^{-1} \tau_2$, then the eigenvalues of $I + \tau_1^{-1} \tau_2$ are $1 + \lambda_j$ ($j = 1, \dots, d$), and therefore from the third equality above the criterion ρ^* becomes:

$$\rho^* = \frac{2^{\frac{d}{2}} \left\{ \prod_{j=1}^d \lambda_j \right\}^{\frac{1}{4}}}{\left\{ \prod_{j=1}^d (1 + \lambda_j) \right\}^{\frac{1}{2}}} = \left\{ \prod_{j=1}^d \frac{2\lambda_j^{\frac{1}{2}}}{1 + \lambda_j} \right\}^{\frac{1}{2}}. \quad (14)$$

(b) Jeffreys' non-informative prior

Jeffreys' non-informative prior for (μ, Σ) , where $\Sigma = r^{-1}$, is of the form:

$$p(\mu, \Sigma) = |\Sigma|^{-(d+1)}. \quad (15)$$

This prior can be taken as a limiting case of the corresponding Normal-Inverse Wishart prior for (μ, Σ) when the matrix $\tau_0 \rightarrow 0$, $\alpha_0 \rightarrow 0$ and $n_0 \rightarrow 0$. With such a prior the measure of affinity ρ^* becomes:

$$\rho^* = \frac{\left\{ |S_1 \parallel S_2| \right\}^{\frac{1}{4}}}{\left| \frac{1}{2} \{S_1 + S_2\} \right|^{\frac{1}{2}}} = 2^{\frac{d}{2}} \frac{|S_1^{-1} S_2|^{\frac{1}{4}}}{|I + S_1^{-1} S_2|^{\frac{1}{2}}} = 2^{\frac{d}{2}} \frac{|S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}}|^{\frac{1}{4}}}{|I + S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}}|^{\frac{1}{2}}}. \quad (16)$$

The above result could also be taken by realising that when the sample size n is very large, $\tau_1 \approx S_1$ and $\tau_2 \approx S_2$. Consequently the result (14) holds also here with λ_j ($j = 1, \dots, d$) being the eigenvalues of $S_1^{-1} S_2$. It must be mentioned here that in this case the number of eigenvalues that are different from zero is: $m = \text{rank}(S_2) = \min\{d, L-1\}$. Since L is generally very large, $m = d$.

4. Optimal partitioning by hyperplanes in R^d

In the following we consider the case of equal groups, i.e. when $k_1 = \dots = k_L = k$. In addition we will work with the result (16) and assume that the data points x_i are 'typical' in R^d , in the meaning that the conditions, in Lemmas 4.1 and 4.3 bellow, that characterise the optimal partition will be satisfied.

Lemma 1. The measure of affinity ρ^* in (16) is maximised when $\lambda_1 = \dots = \lambda_d = \lambda^*$ with λ^* the maximum possible value.

Proof: To maximise (16) we set the constraint $\prod \lambda_j = c$, where c is a positive constant. The relation $S_1 = S_2 + S_3$, with S_1 , S_2 and S_3 positive definite, implies that the eigenvalues of $S_1^{-1} S_2$ are in the interval $(0, 1]$. Thus, c also has to be in the interval $(0, 1]$. Applying the Lagrange multipliers method, we can maximise the quantity ρ^* under the above constraint, and find that ρ^* is maximised when $\lambda_1 = \dots = \lambda_d = \lambda$, with $\lambda = c^{1/d}$. Then the measure of affinity ρ^* becomes equal to $\{2 \sqrt{\lambda} / (1 + \lambda)\}^{d/2}$. Since this is an increasing function of λ in the interval $(0, 1]$, the optimal value for ρ^* is given by

$$\rho^* = \left\{ \frac{2\sqrt{\lambda^*}}{1 + \lambda^*} \right\}^{\frac{d}{2}}, \quad (17)$$

with λ^* the highest possible value.

Lemma 2. The measure of affinity ρ^* in (16) remains the same when we apply the linear transformation $z = S_1^{-1/2} x$ to the original data set X .

Proof: This follows from the last equality in the result (16).

It is interesting to note that Hellinger's distance is invariant to linear transformations, and more generally to any one-to-one transformation, of the data. The transformed data set will be denoted Z .

Lemma 3. The between-group scatter matrix S_2 that corresponds to the optimal partition satisfies the following relations:

$$\begin{aligned} (i) \quad & S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}} = \lambda^* I, \\ (ii) \quad & S_2 = \lambda^* S_1, \\ (iii) \quad & S_1^{-1} S_2 = S_1^{-\frac{1}{2}} S_2 S_1^{-\frac{1}{2}} = \lambda^* I \end{aligned}$$

with $\lambda^* \in (0, 1]$.

Proof: We prove only the first one. Let λ_j ($j = 1, \dots, d$) be the eigenvalues of the matrix $H = S_1^{-1/2} S_2 S_1^{-1/2}$. The argument in Lemma 1 leads again to the conclusion that for the optimal partition we have $\lambda_1 = \dots = \lambda_d = \lambda^*$. Let now Γ be the matrix of eigenvectors of H . Thus, $\Gamma^T H \Gamma = \lambda^* I$ and therefore $H = \lambda^* I$.

Remark. Since the eigenvalues λ_j ($j = 1, \dots, d$) represent variances of group means, it is useful to mention here one of the main conclusions in Defays and Nanopoulos (1992) and in Defays *et al* (2001). In the special case of univariate data it is shown there, that in order to maximise the between-group sum of squares, the extreme points of any group should not be within the range of values of another group. In other words, the group ranges should not overlap each other.

Theorem. The optimal partition that maximises (16) is obtained by partitioning the data set X by hyperplanes perpendicular to the eigenvectors of S_1 . The number of cutting hyperplanes has to be the same for all eigenvectors.

Proof: By Lemma 3 the optimal partition is characterised by the equation: $S_1^{-1/2} S_2 S_1^{-1/2} = \lambda^* I$. Let $G = \{G_1, \dots, G_L\}$ be an arbitrary partition. The eigenvalue λ_j ($= \lambda^*$ for the optimal one) is the variance of the projections of $\bar{z}_l = S_1^{-1/2} \bar{x}_l$ ($l = 1, \dots, L$) onto the j -th axis of the coordinate system. By taking into account the previous remark, the variance λ_j can be maximised when separating the groups by hyperplanes perpendicular to that axis. The larger the number of the hyperplanes that cut the j -th axis perpendicularly, the larger the λ_j . Since for the optimal partition all λ 's have to be equal to each other, the symmetry of the problem implies that the number of cutting hyperplanes has to be the same for all axes.

5. Description of the algorithm

The above result can be implemented in the following way:

Set $p = k/n$ and $\nu = \lceil p^{-1/d} \rceil$, where $\lceil x \rceil$ denotes the integral part of x . Take on each axis of the d -dimensional coordinate system the partition $\{(-\infty, y_1], \dots, (y_{\nu-2}, y_{\nu-1}], (y_{\nu-1}, +\infty)\}$ in ν classes of equal probability under the univariate standard Normal distribution. Form the $L = \nu^d$ hypercubes by taking the $\nu - 1$ hyperplanes perpendicular to each axis at the partition points y_j ($j = 1, \dots, \nu - 1$). (Note that the probability measure of the so derived hypercubes under the d -variate standard Normal distribution is at least p). Evaluate the number m_l of the transformed data points $z_i = S_1^{-1/2} x_i - \bar{x}$, ($i = 1, \dots, n$) that fall within the l -th hypercube for $l = 1, \dots, L$. Then, if $m_l \geq k$ for all l 's we have got the optimal partition. If $m_l \geq k$ for all l 's corresponding to interior hypercubes but there are some l 's corresponding to exterior, i.e. open at $\pm\infty$, hypercubes for which $m_l < k$, then collapse the exterior hypercubes and obtain a suboptimal partition satisfying the condition $m_l \geq k$ for all l 's. If $m_l < k$ for some l 's corresponding to interior hypercubes, then increase p by a small positive amount δ and start again from the beginning until the condition $m_l \geq k$ is satisfied for all l 's. Evaluate the group means \bar{z}_l ($l = 1, \dots, L$).

The set of micro-aggregated data is then formed by the points $\bar{x}_l = \bar{x} + S_1^{1/2} \bar{z}_l$ ($l = 1, \dots, L$) together with their multiplicities m_l .

6. References

- [1] Adam, N. P. and Wortmann, J. C. (1989), 'Security control methods for statistical databases — A comparative study', *ACM Computing Surveys*, 21, pp. 515–556.
- [2] Defays, D. and Nanopoulos, Ph. (1992), 'Panels of enterprises and confidentiality: The small aggregates method', *Proceedings of Statistics Canada Symposium — Design and Analysis of Longitudinal Surveys*, Ottawa, pp. 195–204.
- [3] Defays, D., Nanopoulos, Ph. and Kokolakis, G. (2001), 'Clustering analysis under a cardinality constraint', *Research in Official Statistics* (to appear).
- [4] De Groot, M. H. (1970), *Optimal statistical decisions*, McGraw-Hill, New York.
- [5] Duda, R. O. and Hart, P. E. (1973), *Pattern classification and scene analysis*, Wiley, New York.
- [6] Duncan, G. T. and Lambert, D. (1989), 'The risk of disclosure for microdata', *Journal of Business and Economic Statistics*, 7, pp. 207–217.

- [7] Duncan, G. T. and Pearson, R. W. (1991), 'Enhancing access to microdata while protecting confidentiality: Prospects for the future', *Statistical Science*, 6, pp. 219–239.
- [8] Fienberg, S. E. (1994), 'Conflict between the needs for access to statistical information and demands for confidentiality', *Journal of Official Statistics*, 10, pp. 115–132.
- [9] Fienberg, S. E. and Makov, U. E. (1998), 'Confidentiality, uniqueness and disclosure limitation for categorical data', *Journal of Official Statistics*, 14, pp. 385–398.
- [10] Krzanowski, W. J. (1983), 'Distance between populations using mixed continuous and categorical variables', *Biometrika*, 70, pp. 235–243.
- [11] McGuckin, R. and Nguyen, S. (1988), 'Use of "surrogate files" to conduct economic studies with longitudinal microdata', *Proceedings of the Third Annual Research Conference*, Bureau of the Census, Washington, DC, pp. 193–209.
- [12] McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition, analysis and statistical pattern recognition*, New York.
- [13] Omori, Y. (1999), 'Measuring identification disclosure risk for categorical microdata by posterior population uniqueness', *Statistical Data Protection '98*, Lisbon, Luxembourg, Eurostat.
- [14] Robert, C. P. (1994), *The Bayesian choice: A decision-theoretic motivation*, Springer, New York.
- [15] Samuels, S. M. (1998), 'A Bayesian species-sampling-inspired approach to the uniqueness problem in microdata disclosure-risk assessment', *Journal of Official Statistics*, 14, pp. 373–384.

Using the national longitudinal survey of youth in the United States to study the birth process: A Bayesian approach

Kai Li and Dale J. Poirier

*University of British Columbia, Canada
University of California, Irvine, United States*

Keywords: Bayesian, birth length, birth weight, drinking, gestation, NLSY, prenatal care, simultaneity, smoking, weight gain

Abstract

This paper employs the national longitudinal survey of youth in the United States to study the birth process. We develop a simultaneous equations model with seven endogenous variables: four birth inputs (maternal smoking, maternal drinking, first trimester prenatal care, and maternal weight gain), three birth outputs (gestational age, birth length, and birth weight), and 24 exogenous variables. The estimation is Bayesian. Separate analyses are performed on five different groups: Main Whites, Supplemental Whites, Blacks, Hispanics, and Native Americans. In all groups, we find sizeable correlation between the disturbances in the four input and three output equations and among output disturbances. For gestation, the effect of maternal weight is positive and substantial, while the effect of maternal age is consistently negative and substantial for Main Whites, Blacks, and Hispanics. The effects of smoking, drinking, prenatal care, and weight gain vary in sign and magnitude across the groups. For birth length, male infants are on average longer. The effect of maternal height is noticeable but small in magnitude, and the effect of maternal weight is noticeable only for Main Whites. The effect of smoking is consistently negative, and substantial for Main and Supplemental Whites. The effects of drinking and prenatal care vary across the groups. Both weight gain and gestation have consistently positive effects. For birth weight, male infants are on average heavier except Hispanics. The effect of maternal height is noticeable for Main Whites and Hispanics. The effect of maternal weight is noticeable and consistent across the groups. The effect of smoking is consistently negative, and substantial for Main and Supplemental Whites. The effects of drinking and prenatal care are small and vary across the groups. Weight gain has a small positive effect except Supplemental Whites. The effect of gestation is positive and fairly comparable across the groups.

1. Introduction

This paper draws on two disparate literatures on birth weight (BW): economics and biomedical. The primary distinguishing feature between the two is that the economics literature, unlike the biomedical literature, views many aspects of maternal behaviour, together with birth weight and related birth outputs, as endogenous to the birth process, i.e., they are determined or explained within the system under analysis. Endogenous variables are not even conceptually controlled by the researcher. In contrast exogenous variables are determined outside the system. The distinction tells a lot about the researcher's view of the world, and it is one of the first things to be decided. It has major implications for statistical

modelling, and more importantly, on the questions being asked.

Birth weight is probably the single most important indicator of infant health (e.g. see Institute of Medicine, 1985). It is also a significant predictor of infant mortality, morbidity, coronary heart disease, neurodevelopmental handicaps, and learning disabilities (e.g. see Illsley and Mitchell, 1984 and Poirier, 1998). Birth weight is the result of two processes: (i) the gestational age (G), and (ii) the intrauterine growth rate of the foetus. Gestational age is usually assumed to be approximately two weeks shorter than the period elapsed since last normal menstrual period. In this paper we treat both birth weight and gestation as endogenous in the birth process.

Miller and Merritt (1979) forcefully argue that measurements of crown-heel length, head circumference, mid-arm circumference, and skin folds or other indices of body fat are also important data that should be recorded together with birth weight and gestation for the purpose of predicting future morbidity outcomes. In this paper we work with three birth outputs: birth weight, gestation, and birth length (BL).

Economists view birth weight in the context of a process in which the mother acts as a decision-maker striving to achieve goals subject to constraints. Maternal behaviour provides a variety of inputs into the production of birth outcomes. Such formalism is not the goal here, but the purposeful behaviour of the mother in striving for a healthy infant creates demands for health inputs (e.g. whether to smoke, drink, use drugs, obtain prenatal care, etc.) into a three-output birth production function (BPF). The BPF represents the technical (biological/physiological) relationship between the birth outputs gestation, birth length, and birth weight and the birth inputs smoking (S), drinking alcohol (D), seeking prenatal care in the first trimester (PC), and proper maternal nutrition as measured by weight gain (WG) net of birth weight. The inputs are determined by health input demand functions which describe input choices subject to the constraints the mother faces. The essence of the economist's view is that the mother is attempting to do the best that she can for herself and her child subject to the multiple constraints she faces.

The endogeneity of inputs in the BPF is the important distinguishing statistical feature between the economists' models and those of other social scientists and epidemiologists. It builds on the seminal work of Grossman (1972) who introduces the idea of a health production function relating health outcomes, via physiological or biological processes, to health inputs chosen by the individual. Such inputs are generally desired, not because they directly provide utility, but because they have an instrumental role to play in producing goods (e.g. health) that are valued directly.

2. Data

The statistical window to be employed in this paper is quite ambitious compared to counterparts in the biomedical literature on birth weight, and so we employ a very rich data set commonly used by social scientists, the National Longitudinal Survey of Youth (NLSY) in the United States for implementation.

The NLSY is an ongoing study of 12 686 young men and women aged 14 to 21 as of 1 January 1979. Over 90 % of these respondents have participated in an annual personal interview, approximately one hour in length, since 1979. Individuals are followed after leaving their baseline household. There is relatively little attrition.

The data for this paper are drawn from the NLSY Merged Child-Mother file (NLSCM) for 1994 (CD-ROM). Where necessary, additional variables are constructed using the data from the NLSY main file for 1994 (CD-ROM). The price indices on cigarette, alcohol, medical services and food are obtained from the consumer price index database of the Bureau of Labour Statistics.

The NLSCM contains data for each child born to a woman in the original NLSY survey, as well as a selection of variables from the NLSY. Blacks, Hispanics, and the poor were over sampled in the NLSY. Of 6 283 women who began the survey in 1979, 4 599 had given birth to 10 042 children by 1994. Our sample of births is obtained by imposing the constraints that the birth order of the child (G0005800) is 1 and the birth year (G0005700) is after and including 1979. We also drop observations if they miss exogenous variables such as income or income exceeds USD 100 000, AFQT scores, etc. Finally, we drop observations if they miss endogenous variables.

In this paper, we analyse racial/ethnic groups separately. Racial/ethnic groups are defined by the mother's self-reported identification. We examine five racial/ethnic groups: Main Whites, Supplemental Whites, Blacks, Hispanics, and Native Americans.

We choose to analyse only singleton first-born live births, leaving aside sample selection problems arising from parity considerations and abortions. There were 3 648 live singleton first births to White, Black, Hispanic, and Native American women between 1979 and 1994 in the NLSY. We dropped 221 births to women in the military and 28 to women no longer living in the United States. Births to women in the military are sufficiently different from births in the civilian population, so we do not want to contaminate our much larger number of civilian births. The births to emigrants were dropped because of their small sample sizes and our expectation that they should not be combined with our other data. This left 3 399 observations for our target sample.

As in most empirical studies, missing observations are a reality. 35.4 % of our target sample was dropped due to missing observations on at least 1 of our 24 exogenous variables — household income being the primary culprit. More disturbingly, another 6.9 % of the observations have exogenous variable data, but are missing data on one or more of the seven endogenous variables. This leaves a total of 1 962 observations with complete data (57.7 % of our target sample). Missing data is more of a problem for the Black, Hispanic, and Native American samples than for the White samples. Also, the Supplemental White sample has a slightly more severe missing data problem than the Main White sample. For more details, see Li and Poirier (2000, Table 1).

Our choice of the 24 exogenous (conditioning) variables is guided by the existing literature. Variables x_1 is the intercept term. Variables $x_2 - x_6$ cover basic physical characteristics (the gender of the infant, the age and size of the mother) which we expect to be very important in the birth output equations ⁽¹⁾. Variables $x_7 - x_{12}$ capture regional and temporal effects plus the intelligence and family income of the mother ⁽²⁾. Variables $x_{13} - x_{25}$ capture health insurance status and a variety of socioeconomic measures of the mother's family background. Variables $x_7 - x_{25}$ are risk factors that causally are quite far removed from the biological event of low birth weight. We expect these variables to be important in the input equations, but not in the biologically-based output equations.

One variable notably missing is the marital status of the mother or whether she is living with the father. Clearly such measures are endogenous, and furthermore, reflect an endogenous decision by the father as well. Just as we are not modelling fertility (see footnote 1), we are not trying to model the marriage/cohabit decision. Implicitly we are conditioning on the decisions to get pregnant and not to have an abortion. We do not feel a latent distribution, say, of potential birth weight for infants not conceived, or conceived but aborted, is of great interest. We are no more willing to condition on the marriage/cohabit decision than to condition on the decision to smoke. So we have marginalised out the marriage/cohabit decision from our model. Note, however, the presence of the father can be reflected in variables such as the household income (x_{12}) and the number of adults in household (x_{15}). Also note that there are many missing observations for whether the father is present in the household.

3. Modelling

Following the strategy outlined in Poirier (1995, Chapter 10), we choose a highly over-identified specification for our maintained hypothesis H_* , and a less restricted specification H_A as an alternative hypothesis that we expect will not lead to rejecting H_* . Our prior reflects this viewpoint. In Section 4.1 we test these over-identifying restrictions.

Our model specification is the same as used in Li and Poirier (2001). Our distribution of interest, for singleton first-born live births, is the joint distribution of four birth inputs (smoking, drinking, prenatal care, and weight gain) and three birth outputs (gestation, birth length, and birth weight), given the exogenous variables x . We choose a fairly large 155-dimensional parametric window to model this seven-dimensional conditional distribution of endogenous variables z .

Consider a sample of T independent singleton first-born live births indexed by the subscript i . Let $[S_i^*, D_i^*, PC_i^*]'$ ($i = 1, 2, \dots, T$) denote latent variables underlying the

⁽¹⁾ We are not trying to explain fertility, and so we are not trying to explain the mother's pregnancy. Hence, variables like maternal age (x_6) are properly treated as exogenous in our analysis.

⁽²⁾ None of the 81 Native Americans in Li and Poirier (1999) lived in the Northeast. Therefore, the specification for Native American mothers contains only 23 exogenous variables plus the constant term in the S , D , PC , and WG equations.

binary birth inputs $[S_i, D_i, PC_i]' = [\mathbf{1}(S_i^*), \mathbf{1}(D_i^*), \mathbf{1}(PC_i^*)]'$ ($i = 1, 2, \dots, T$), where $\mathbf{1}(\bullet)$ denotes an indicator function which equals unity if the argument is positive and equals zero otherwise. For estimation, we partition the endogenous variables into inputs z_{i1} and outputs z_{i2} : $z_{i1}^* = [S_i^*, D_i^*, PC_i^*, WG_i^*]'$, $z_{i1} = [S_i, D_i, PC_i, WG_i]'$, $z_{i2} = [G_i, BL_i, BW_i]'$ ($i = 1, 2, \dots, T$). Let x_i ($i = 1, 2, \dots, T$) denote $K \times 1$ vectors of exogenous variables.

Suppose the four inputs are generated from the following specification

$$z_{i1}^* = \Delta_1' x_i + \varepsilon_{i1}, \quad (1)$$

where $\Delta_1 = [\Delta_S, \Delta_D, \Delta_{PC}, \Delta_{WG}]$ is $K \times 4$. Also suppose the three birth outputs are related to $z_{i1} = [S_i, D_i, PC_i, WG_i]'$ as follows:

$$z_{i2}' \Gamma_2 = z_{i1}' \Gamma_1 + x_i' \Delta_2 + \varepsilon_{i2}', \quad (2)$$

where $\varepsilon_i = [\varepsilon_{i1}, \varepsilon_{i2}]' \mid x_i \sim \text{i.i.d. } N_7(0_7, \Sigma)$ ($i = 1, 2, \dots, T$), Γ_2 is nonsingular

$$\Gamma_1 = \begin{bmatrix} \gamma_{S,G} & \gamma_{S,BL} & \gamma_{S,BW} \\ \gamma_{D,G} & \gamma_{D,BL} & \gamma_{D,BW} \\ \gamma_{PC,G} & \gamma_{PC,BL} & \gamma_{PC,BW} \\ \gamma_{WG,G} & \gamma_{WG,BL} & \gamma_{WG,BW} \end{bmatrix}, \quad (3)$$

$$\Gamma_2 = \begin{bmatrix} 1 & -\gamma_{G,BL} & -\gamma_{G,BW} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

$$\Delta_2 = \begin{bmatrix} \delta_G & \delta_{BL} & \delta_{BW} \\ \delta_{5,G} & \delta_{5,BL} & 0 \\ \delta_{6,G} & 0 & \delta_{6,BW} \\ \Delta_{*,G} & \Delta_{*,BL} & \Delta_{*,BW} \\ 0_{13} & 0_{13} & 0_{13} \end{bmatrix}, \quad (5)$$

where $\delta_j = [\delta_{1,j}, \delta_{2,j}, \delta_{3,j}, \delta_{4,j}]'$, and $\Delta_{*,j} = [\delta_{7,j}, \dots, \delta_{12,j}]'$, ($j = G, BL, BW$). The coefficients in $\Delta_{*,j}$ ($j = G, BL, BW$) are set to zero under our maintained specification. Finally, $\Sigma = [\Sigma_{ij}]$ ($i, j = 1, 2$) is partitioned into the four birth inputs and the three birth

outputs.

The specification in equations (1) - (5) warrants a few comments. It reflects a view of the world in which reduced form (equation (1)) is postulated for the four inputs (smoking, drinking, prenatal care, and weight gain), and then a triangular view (equations (2) and (4)) of the three outputs (gestation, birth length, and birth weight) is postulated in which gestation is determined based on the four inputs, and then birth length and birth weight are jointly determined as functions of the four inputs and gestation. The three output equations are identified by zero restrictions on maternal weight (x_6) in the birth length equation, and on maternal height (x_5) in the birth weight equation. The model is not recursive because Σ is permitted to be non-diagonal. The model is non-linear because of the jointly determined dummy endogenous variables (smoking, drinking, and prenatal care). The specification of numerous zero restrictions on Δ_2 in equation (5) ensures that the order condition for identification is satisfied.

Our prior is proper, but moderately diffuse. We use the same prior for all racial/ethnic groups. The estimation of our model extends the work by Chib and Greenberg (1998) and Li (1998), and is described in Li and Poirier (2000, Appendices A.3-A.4). To give a quick, visual indication of the posterior mass around the means, we indicate the relative size of the posterior mean to the posterior standard deviation by the border of the table cell as described in Table 1.

4. Empirical results

4.1. Evidence of structure

We investigate whether our output equations reflect a biological structure in three related ways. First, the logarithmic Bayes factor in favour of our maintained specification H^* : $\Delta^*_{\cdot G} = \Delta^*_{\cdot BL} = \Delta^*_{\cdot BW} = 0_6$ versus the alternative H_A : $\Delta^*_{\cdot G} 0_6$ or $\Delta^*_{\cdot BL} 0_6$ or $\Delta^*_{\cdot BW} 0_6$ is overwhelming for all groups (Li and Poirier, 2000, Table 10). Second, the predictive densities for all endogenous variables differ little across H^* and H_A (Li and Poirier, 2000, Table 11). Third, under H_A the six additional variables $x_7 - x_{12}$ add relatively little to the three output equations (Li and Poirier, 2000, Table 12). Because of these results, subsequent results are conditioned upon H^* .

4.2. System results

Our treatment of simultaneity, in contrast to most of the biomedical literature, is a distinguishing feature of our model. While our window imposes triangularity, it does not impose a full recursive specification. Our prior for Σ is centred over a diagonal matrix (supporting the use of single-equation methods), the need for simultaneous equations techniques is apparent in our posterior results (Li and Poirier, 2000, Table 13). Although the correlations between input disturbances and the birth weight disturbance are fairly small for most groups, this should not be interpreted as justifying simply running a regression for the

birth weight equation. The correlation between the disturbances in the gestation and birth weight equations is sizeable. Indeed, as Li and Poirier (2000, Section 4.6) note, the ordinary least squares results for the birth weight equation of Main Whites are substantially different from our posterior results.

4.3. Input equations

Our interest in the parameters of the input equations is minimal compared to the output equations, and so we devote less attention to them. Tables 2a-2d contain the posterior (group-specific) and prior means and standard deviations for the elements of Δ_1 under our default prior. Some results are not very surprising. For example, the posterior mass for the coefficient of the AFQT score variable in the smoking equation is negative and large relative to its standard deviation for all groups, except Hispanics. On the other hand, the coefficient of the AFQT score variable in the drinking equation is positive and large relative to its standard deviation for both the Main White and Supplemental White groups. Note that in both cases the priors for the coefficients are located away from zero, but in the latter case, the prior posterior locations differ in sign. Clearly our priors are not dominating the data.

The price indices do not appear to serve very well as instruments in any of the input equations for any of the groups. But most other variables among $x_7 - x_{25}$ have substantial posterior mass away from zero in some equations for every group suggesting they satisfy at least one requirement of a legitimate instrumental variable for the output equations.

4.4. Output equations

The output equations are of prime importance. They describe how birth inputs together with the biological size of the mother are transformed into birth outputs describing the physical characteristics of the infant. We discuss each of the three equations in turn, presenting posterior results under the default prior. When discussing maternal height and weight we take into account both their effects through body mass index [$\text{BMI} = \text{weight in kg}/(\text{height in m})^2$] and their linear effects. The posterior means and standard deviations of the partial derivatives of the exogenous variable effects of maternal height and weight are reported for each output equation.

The posterior results for the gestation equation are reported in Table 3. The pictures regarding the effects of exogenous variables differ somewhat across groups. Although BMI, maternal height, and maternal weight do not appear to matter much individually for Main Whites, Blacks, and Hispanics, the net marginal effect of maternal weight is substantial and similar across the groups. In contrast, the same three variables appear to have separate effects for Supplemental Whites and Native Americans, which yield no net effects for Supplemental Whites, and a negative net effect of maternal height for Native Americans. The posterior effect of being male is only noticeable for Main Whites. The posterior effect of maternal age is consistently negative and noticeably shifted away from the origin for Main Whites, Blacks, and Hispanics.

The posterior mean effects of the three endogenous binary inputs vary in sign and magnitude across the groups. Of the 15 (3H5) cases, a posterior mean is more than twice its standard deviation only twice. There is more consistency in the effects of weight gain on gestation across the groups, and in most cases the effects of weight gain are small.

The posterior results for the birth length equation are reported in Table 4. Similar pictures emerge regarding the effects of exogenous variables across the groups, except for the large standard deviations in the small sample of Native Americans. Clearly, the birth lengths of male infants are on average longer. The net marginal effect of maternal height on birth length is noticeable but small in magnitude. The net marginal effect of maternal weight is noticeable only for Main Whites.

The posterior mean effects of the three endogenous binary inputs are more similar across the groups in the birth length equation than they are in the gestation equation. The posterior mean effect of smoking on birth length is consistently negative across the groups, and larger than its standard deviation for Main and Supplemental Whites. The posterior mean effects of drinking and prenatal care on birth length vary across the groups. Both weight gain and gestation have positive mean effects on birth length, which are quite consistent across the groups.

The posterior results for the birth weight equation are reported in Table 5. Similar pictures emerge regarding the effects of exogenous variables across the groups, except for the large standard deviations in the small sample of Native Americans. Clearly, male infants are on average heavier, except in the case of Hispanics. The net marginal effect of maternal height on birth weight is noticeable for Main Whites and Hispanics. The net marginal effect of maternal weight on birth weight is noticeable, and consistent across the groups.

Like in the birth length equation, the posterior mean effects of the three endogenous binary inputs are more similar across the groups in the birth weight equation than they are in the gestation equation. The posterior mean effect of smoking on birth weight is consistently negative across the groups, and larger than its standard deviation for Main and Supplemental Whites. The posterior mean effects of drinking and prenatal care on birth weight vary across the groups, and are generally small. Weight gain has a small positive mean effect on birth weight for all groups except Supplemental Whites in which case it is negative and larger in absolute value than its posterior standard deviation. The posterior mean effect of gestation on birth weight is positive and fairly comparable across the groups.

4.5. Prediction

Given out-of-sample values of the exogenous variables x , the predictive density for the out-of-sample of \tilde{z}^* =

$$[\tilde{z}_1^*, \tilde{z}_2^*]' = [\tilde{S}^*, \tilde{D}^*, \tilde{PC}^*, \tilde{WG}, \tilde{G}, \tilde{BL}, \tilde{BW}]' \text{ is}$$

$$\begin{aligned} f(\tilde{z}_1^*, \tilde{z}_2 | \tilde{x}, Z, X) &= \int_{\Theta} f(\tilde{z}_1^*, \tilde{z}_2 | \theta) f(\theta | Z, X) d\theta \\ &= \int_{\Theta} \phi_4(\tilde{z}_1^* | \Delta_1' \tilde{x}, \Sigma_{11}) \phi_3(\tilde{z}_2 | \tilde{\mu}_2 + \Sigma_{12}' \Sigma_{11}^{-1} (\tilde{z}_1^* - \Delta_1' \tilde{x}), \Sigma_{2|1}) f(\theta | Z, X) d\theta, \end{aligned} \quad (6)$$

where $\tilde{\mu}_2 = \tilde{\mu}_2(\tilde{z}_1, \tilde{G}, \tilde{x}, \beta) = \tilde{W}_2 \beta, \Sigma_{2|1} = \Sigma_{22} - \Sigma_{12}' \Sigma_{11}^{-1} \Sigma_{12}$, and

$$\tilde{W}_2 = \begin{bmatrix} \tilde{z}_1' & \tilde{x}^* & \tilde{x}_5 & \tilde{x}_6 & 0 & 0_4' & 0_4' & 0 & 0 & 0_4' & 0_4' & 0 \\ 0_4' & 0_4' & 0 & 0 & \tilde{G} & \tilde{z}_1' & \tilde{x}^* & \tilde{x}_5 & 0 & 0_4' & 0_4' & 0 \\ 0_4' & 0_4' & 0 & 0 & 0 & 0_4' & 0_4' & 0 & \tilde{G} & \tilde{z}_1' & \tilde{x}^* & \tilde{x}_6 \end{bmatrix}, \quad (7)$$

with $\tilde{x}_i^* = [\tilde{x}_{i1}, \tilde{x}_{i2}, \tilde{x}_{i3}, \tilde{x}_{i4}]'$, and

$\tilde{z}_1 = [\tilde{S}, \tilde{D}, \tilde{PC}, \tilde{WG}]' = [1(\tilde{z}_1^*), 1(\tilde{z}_2^*), 1(\tilde{z}_3^*), \tilde{WG}]'$. We will concentrate on the predictive distribution for birth outputs, obtained from equation (6) by integrating out inputs:

$$\begin{aligned} f(\tilde{z}_2 | \tilde{x}, Z, X) &= \int_{\Theta} \left[\int_{\mathbb{R}^4} f(\tilde{z}_1^*, \tilde{z}_2 | \theta) d\tilde{z}_1^* \right] f(\theta | Z, X) d\theta \\ &= \int_{\Theta} \left[\int_{\mathbb{R}^4} \phi_4(\tilde{z}_1^* | \Delta_1' \tilde{x}, \Sigma_{11}) \phi_3(\tilde{z}_2 | \tilde{\mu}_2 + \Sigma_{12}' \Sigma_{11}^{-1} [\tilde{z}_1^* - \Delta_1' \tilde{x}], \Sigma_{2|1}) d\tilde{z}_1^* \right] f(\theta | Z, X) d\theta \\ &= E_{\theta | X, Z} \left[E_{\tilde{z}_1^* | \tilde{x}, \theta} \left(\phi_3(\tilde{z}_2 | \tilde{\mu}_2 + \Sigma_{12}' \Sigma_{11}^{-1} [\tilde{z}_1^* - \Delta_1' \tilde{x}], \Sigma_{2|1}) \right) \right]. \end{aligned} \quad (8)$$

The univariate predictive densities for gestation, birth length, and birth weight are shown in Figures 1-3, respectively. These figures depict the univariate predictive output densities for each group and the very diffuse prior predictive density embodying only the informative prior and no data.

5. Discussion

It is well acknowledged that birth weight is probably the single most important indicator of infant health. In this paper, we focus on explaining the birth outcomes such as gestation, birth length and birth weight using a simultaneous equations approach. On the other hand, the more interesting and ultimately relevant question to ask, from a society viewpoint, is what factors affect children's attainment later in life. Our modelling framework turns out to be quite useful in answering questions like this. We conjecture that birth weight and related birth measurements are the intervening variables in explaining children's development later in life, and we plan to investigate further in future work.

6. Acknowledgements

The authors gratefully acknowledge the research support of the Social Sciences and Humanities Research Council of Canada, and helpful comments on earlier versions by Sid Chib, John Geweke, Gary Koop, and conference participants at the Sixth World Meeting of the International Society for Bayesian Analysis at Crete, and the 2000 World Congress of the Econometric Society at Seattle. The usual disclaimer applies.

7. Tables

Table 1: Notational conventions in subsequent tables




	Absolute value of mean between one and two standard deviations
	Absolute value of mean between two and three standard deviations
	Absolute value of mean more than three standard deviations
bold	Standard deviation equal to zero

Table 2a: Posterior means (standard deviations) of the effects of maternal biological characteristics on birth inputs under H*

	S	D	PC	WG	S	D	PC	WG
	Main White				Black			
x ₁ Intercept	-.2277 (.1111)	.2007 (.1006)	.8230 (.1225)	11.77 (.4960)	-.6422 (.1971)	-.2436 (.1801)	.7437 (.1863)	11.28 (1.015)
x ₂ Male child	-.0336 (.0851)	.0282 (.0837)	.2040 (.0982)	.1780 (.3651)	-.2098 (.1346)	-.0685 (.1252)	-.0234 (.1313)	.6083 (.6658)
x ₃ Mother's age — 23yrs.	.0083 (.0222)	.0581 (.0203)	.0530 (.0269)	-.0854 (.0994)	.0012 (.0390)	.0674 (.0342)	.0848 (.0373)	.1366 (.1880)
x ₄ Body mass index — 24	.1161 (.1467)	-.1351 (.1343)	-.0673 (.1660)	-.0833 (.6882)	.1970 (.1510)	-.0213 (.1323)	.1758 (.1337)	-.4462 (.7060)
x ₅ Maternal height — 162cm	.0359 (.0423)	-.0208 (.0394)	-.0208 (.0491)	.0417 (.1992)	.0746 (.0468)	.0001 (.0410)	.0388 (.0407)	-.1251 (.2199)
x ₆ Maternal weight — 63kg	-.0390 (.0542)	.0510 (.0496)	.0180 (.0612)	.0671 (.2534)	-.0731 (.0561)	.0161 (.0486)	-.0642 (.0484)	.2608 (.2596)
	Supplemental White				Native American			
x ₁ Intercept	-.3272 (.2052)	.0638 (.2001)	.5734 (.2308)	10.20 (1.017)	.1204 (.3035)	.0833 (.3080)	.3478 (.3435)	9.990 (1.851)
x ₂ Male child	.2504 (.1528)	.0712 (.1533)	.0522 (.1850)	.9329 (.7692)	.3487 (.2487)	-.0095 (.2449)	.7742 (.2876)	-.1443 (1.489)
x ₃ Mother's age — 23yrs.	-.0631 (.0425)	.0013 (.0396)	.1748 (.0538)	-.3841 (.1953)	-.0736 (.0848)	.2202 (.0924)	.1547 (.1078)	.2917 (.5083)
x ₄ Body mass index — 24	-.3169 (.2379)	-.1433 (.2299)	.4587 (.3329)	3.141 (1.098)	.1911 (.4783)	.8289 (.4649)	.9296 (.4853)	1.976 (2.638)
x ₅ Maternal height — 162cm	-.1017 (.0676)	-.0525 (.0660)	.1157 (.0917)	.8819 (.3130)	.0613 (.1481)	.2653 (.1465)	.2531 (.1542)	.5451 (.8290)
x ₆ Maternal weight — 63kg	.1188 (.0883)	.0455 (.0855)	-.1584 (.1212)	-1.087 (.4093)	-.0796 (.1804)	-.3246 (.1750)	-.3544 (.1850)	-.6503 (.9914)

	S	D	PC	WG
Hispanic				
x ₁ Intercept	-.0258 (.2656)	.3534 (.2444)	.3769 (.2494)	11.00 (1.343)
x ₂ Male child	.0339 (.1644)	-.0144 (.1374)	.1402 (.1434)	.3779 (.7260)
x ₃ Mother's age — 23yrs.	-.0361 (.0483)	.0131 (.0381)	.0502 (.0408)	-.4414 (.2002)
x ₄ Body mass index — 24	-.0212 (.2697)	.0183 (.2200)	.2392 (.2424)	.9538 (1.204)
x ₅ Maternal height — 162cm	.0044 (.0821)	.0130 (.0671)	.0709 (.0720)	.3861 (.3657)
x ₆ Maternal weight — 63kg	.0261 (.1041)	-.0137 (.0854)	-.0920 (.0936)	-.2943 (.4685)
Prior				
x ₁ Intercept	.0000 (.4400)	.0000 (.4400)	.0000 (.4400)	10.00 (2.640)
x ₂ Male child	.0000 (.4400)	.0000 (.4400)	.0000 (.4400)	.0000 (2.640)
x ₃ Mother's age — 23yrs.	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
x ₄ Body mass index — 24	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
x ₅ Maternal height — 162cm	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
x ₆ Maternal weight — 63kg	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)

Table 2b: Posterior Means (Standard Deviations) of the Effects of Regional, Temporal, Intelligence and Income on Birth Inputs Under H*

		S	D	PC	WG	S	D	PC	WG
		Main White				Black			
x ₇	Northeast	-.0641 (.1401)	-.0591 (.1277)	.2358 (.1680)	.4640 (.6803)	.3781 (.2334)	.1437 (.2106)	-.0929 (.2246)	.3435 (1.239)
x ₈	South	-.0878 (.1106)	-.4158 (.1046)	-.0627 (.1311)	.4300 (.5148)	-.3808 (.1780)	-.3989 (.1575)	-.0659 (.1743)	-.3336 (.9158)
x ₉	West	.0542 (.1227)	-.1613 (.1130)	-.1035 (.1405)	.6132 (.5843)	.1991 (.2553)	.0303 (.2403)	.1536 (.2708)	-1.155 (1.408)
x ₁₀	Calendar Time — (19)85	.1536 (.0945)	.0345 (.0786)	-.0990 (.1066)	-.3975 (.4104)	.0948 (.1659)	-.0647 (.1383)	.0613 (.1491)	-1.345 (.8305)
x ₁₁	(AFQT score / mean of same age) -1	-.5445 (.0781)	.2555 (.0717)	-.0363 (.0884)	-1.407 (.3417)	-.1936 (.1723)	-.0163 (.1474)	.1097 (.1579)	-.1578 (.8647)
x ₁₂	Household income in \$1 000 — 25	-.0028 (.0031)	.0079 (.0028)	.0103 (.0036)	.0060 (.0123)	-.0042 (.0057)	.0032 (.0049)	.0018 (.0056)	-.0225 (.0288)
		Supplemental White				Native American			
x ₇	Northeast	.4713 (.2296)	.4313 (.2203)	.0377 (.2743)	1.031 (1.083)	.0000 (.0000)	.0000 (.0000)	.0000 (.0000)	.0000 (.0000)
x ₈	South	.1803 (.2063)	-.0386 (.1953)	.1396 (.2451)	.2917 (.9571)	-.4757 (.3275)	-.1877 (.3337)	-.1193 (.3915)	1.792 (2.008)
x ₉	West	.0458 (.2373)	.4093 (.2256)	.0978 (.2762)	2.436 (1.111)	.2280 (.3929)	.0602 (.4206)	.1241 (.4813)	-.3904 (2.409)
x ₁₀	Calendar Time — (19)85	.1915 (.3047)	-.1544 (.2920)	.2762 (.3225)	.9472 (1.573)	.3720 (.3004)	-.2141 (.2963)	-.5477 (.3476)	-1.864 (1.836)
x ₁₁	(AFQT score / mean of same age) -1	-.2285 (.1574)	.4173 (.1504)	.2574 (.2039)	.7960 (.7503)	-.5711 (.2575)	-.2343 (.2709)	.2118 (.3004)	.4115 (1.501)
x ₁₂	Household income in \$1 000 — 25	-.0100 (.0076)	.0060 (.0070)	.0086 (.0097)	.0090 (.0345)	-.0077 (.0130)	.0158 (.0138)	-.0034 (.0142)	-.0524 (.0765)

	S	D	PC	WG
	Hispanic			
x ₁ Intercept	<div><div>-0.2193</div><div>(.3276)</div></div>	<div><div>-0.2729</div><div>(.2946)</div></div>	<div><div>.4148</div><div>(.3075)</div></div>	<div><div>2.058</div><div>(1.618)</div></div>
x ₂ Male child	<div><div>-.5856</div><div>(.2922)</div></div>	<div><div>-.3954</div><div>(.2478)</div></div>	<div><div>.2561</div><div>(.2645)</div></div>	<div><div>.5980</div><div>(1.399)</div></div>
x ₃ Mother's age — 23yrs.	<div><div>-.6129</div><div>(.2758)</div></div>	<div><div>-.2699</div><div>(.2384)</div></div>	<div><div>.1136</div><div>(.2423)</div></div>	<div><div>.7121</div><div>(1.324)</div></div>
x ₄ Body mass index — 24	<div><div>-.4829</div><div>(.2048)</div></div>	<div><div>.0141</div><div>(.1579)</div></div>	<div><div>-.0095</div><div>(.1781)</div></div>	<div><div>.8990</div><div>(.8803)</div></div>
x ₅ Maternal height — 162cm	<div><div>-.0628</div><div>(.1804)</div></div>	<div><div>-.0243</div><div>(.1521)</div></div>	<div><div>.1273</div><div>(.1538)</div></div>	<div><div>1.641</div><div>(.8220)</div></div>
x ₆ Maternal weight — 63kg	<div><div>-.0072</div><div>(.0070)</div></div>	<div><div>-.0009</div><div>(.0051)</div></div>	<div><div>.0008</div><div>(.0055)</div></div>	<div><div>.0117</div><div>(.0287)</div></div>
	Prior			
x ₁ Intercept	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(3.734)</div></div>
x ₂ Male child	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(3.734)</div></div>
x ₃ Mother's age — 23yrs.	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(.6223)</div></div>	<div><div>.0000</div><div>(3.734)</div></div>
x ₄ Body mass index — 24	<div><div>-.2000</div><div>(3.000)</div></div>	<div><div>-.2000</div><div>(3.000)</div></div>	<div><div>.2000</div><div>(3.000)</div></div>	<div><div>.2000</div><div>(18.00)</div></div>
x ₅ Maternal height — 162cm	<div><div>-1.000</div><div>(.4400)</div></div>	<div><div>-1.000</div><div>(.4400)</div></div>	<div><div>1.000</div><div>(.4400)</div></div>	<div><div>1.000</div><div>(2.640)</div></div>
x ₆ Maternal weight — 63kg	<div><div>.0000</div><div>(3.000)</div></div>	<div><div>.0000</div><div>(3.000)</div></div>	<div><div>.0000</div><div>(3.000)</div></div>	<div><div>.0000</div><div>(18.00)</div></div>

Table 2c: Posterior Means (Standard Deviations) of the Effects of Insurance and Socioeconomic Characteristics on Birth Inputs Under H*

	S	D	PC	WG	S	D	PC	WG
	Main White				Black			
X13 No health insurance available	.2664 (.1257)	.0025 (.1112)	-.1330 (.1425)	.3154 (.5580)	-.2805 (.2349)	.0502 (.2042)	-.1130 (.2124)	-1.171 (1.170)
X14 Missing health insurance availability	-.1564 (.1337)	.0121 (.1247)	.0936 (.1564)	.1454 (.6595)	.2619 (.2526)	.3461 (.2181)	.1666 (.2141)	.8354 (1.196)
X15 Number of adults in household — 2	.0290 (.0580)	-.0205 (.0545)	-.1201 (.0665)	-.3854 (.2788)	.0195 (.0591)	-.0447 (.0542)	.0490 (.0606)	.1798 (.3035)
X16 Number of quarters worked last year — 3	.0110 (.0402)	.0489 (.0348)	.0143 (.0466)	.0119 (.1816)	-.0123 (.0590)	.0713 (.0513)	.0409 (.0564)	.3348 (.2881)
X17 Number of maternal siblings — 4	.0015 (.0225)	-.0134 (.0210)	-.0274 (.0269)	.0358 (.1036)	.0261 (.0247)	-.0041 (.0216)	.0412 (.0233)	.0878 (.1205)
X18 Grandmother's education — 12yrs.	-.0042 (.0227)	-.0116 (.0197)	.0139 (.0263)	.0665 (.0988)	.0191 (.0343)	.1025 (.0309)	-.0012 (.0309)	.3446 (.1623)
	Supplemental White				Native American			
X13 No health insurance available	.0342 (.2212)	-.3887 (.2057)	.3010 (.2567)	3.343 (1.074)	-.0442 (.2949)	.0090 (.2956)	.1658 (.3272)	-1.106 (1.783)
X14 Missing health insurance availability	-.0305 (.2377)	.3292 (.2226)	-.3041 (.2815)	-1.315 (1.125)	-.4918 (.3892)	-.5810 (.3966)	.0944 (.4338)	2.462 (2.271)
X15 Number of adults in household — 2	.1346 (.1257)	.1741 (.1145)	-.0813 (.1497)	.0995 (.5658)	-.3236 (.2002)	.1267 (.1812)	.5260 (.2928)	-.1144 (1.010)
X16 Number of quarters worked last year — 3	-.1101 (.0697)	-.0224 (.0656)	-.2206 (.0927)	.3056 (.3206)	.0207 (.1632)	-.3350 (.1662)	-.2693 (.2040)	-.9104 (.9445)
X17 Number of maternal siblings — 4	.0267 (.0386)	-.0178 (.0344)	-.0626 (.0443)	-.0917 (.1666)	-.0004 (.0744)	.0784 (.0817)	-.0009 (.0979)	.1914 (.4590)
X18 Grandmother's education — 12yrs.	.0358 (.0369)	.0404 (.0356)	-.1554 (.0526)	-.0347 (.1732)	-.0608 (.0757)	.0543 (.0844)	.0327 (.0901)	.4149 (.4874)

		S	D	PC	WG
			Hispanic		
X ₁₃	No health insurance available	-0.4102 (.2523)	-.0203 (.1920)	-.0524 (.1955)	.9718 (1.100)
X ₁₄	Missing health insurance availability	.0136 (.2665)	.0145 (.2068)	.3449 (.2084)	.4599 (1.136)
X ₁₅	Number of adults in household — 2	-.0302 (.0709)	-.0394 (.0601)	-.0339 (.0589)	.1139 (.3145)
X ₁₆	Number of quarters worked last year — 3	-.1380 (.0748)	.0980 (.0597)	.1092 (.0598)	.3822 (.3154)
X ₁₇	Number of maternal siblings — 4	.0154 (.0380)	-.0077 (.0291)	-.0699 (.0303)	-.0321 (.1530)
X ₁₈	Grandmother's education — 12yrs.	.1108 (.0285)	.0387 (.0231)	-.0120 (.0219)	.1081 (.1105)
			Prior		
X ₁₃	No health insurance available	.0000 (.4400)	.0000 (.4400)	.0000 (.4400)	.0000 (2.640)
X ₁₄	Missing health insurance availability	.0000 (.6223)	.0000 (.6223)	.0000 (.6223)	.0000 (3.734)
X ₁₅	Number of adults in household — 2	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
X ₁₆	Number of quarters worked last year — 3	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
X ₁₇	Number of maternal siblings — 4	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (18.00)
X ₁₈	Grandmother's education — 12yrs.	-.5000 (.4400)	-.5000 (.4400)	.5000 (.4400)	.5000 (2.640)

Table 2d: Posterior Means (Standard Deviations) of the Effects of Age 14 Status and Prices on Birth Inputs Under H*

		S	D	PC	WG	S	D	PC	WG
		Main White				Black			
x ₁₉	Not on time in school at age 14	.5649 (.1810)	.1810 (.1739)	-.1048 (.1878)	-.1295 (.8600)	.2720 (.2119)	.0382 (.2011)	-.0640 (.2037)	-2.335 (1.146)
x ₂₀	Non-urban at age 14	-.1090 (.1023)	-.0875 (.0841)	.2437 (.1246)	-.2832 (.4522)	-.2434 (.1827)	.0088 (.1611)	.1179 (.1665)	-.2837 (.8985)
x ₂₁	No employed males in household at age 14	-.0635 (.1205)	-.0675 (.1043)	-.0618 (.1356)	.4368 (.5550)	.0466 (.1408)	-.0423 (.1266)	.1890 (.1341)	.5997 (.7171)
x ₂₂	Cigarette price index	-1.545 (.8658)	-1.267 (.7760)	.1862 (.9027)	3.193 (3.863)	.6742 (1.275)	.5820 (1.142)	-1.303 (1.221)	-2.321 (6.840)
x ₂₃	Alcohol price index	-.9823 (1.851)	.1858 (1.705)	-.3937 (1.915)	.8020 (9.001)	-.5599 (2.368)	-.0187 (2.134)	.2852 (2.228)	9.216 (12.95)
x ₂₄	Medical services price index	-.0825 (1.713)	.8312 (1.568)	.1271 (1.750)	-1.284 (7.916)	-1.339 (2.235)	-.8529 (2.050)	.0512 (2.156)	12.47 (12.77)
x ₂₅	Food price index	.3295 (1.723)	-1.337 (1.658)	.7747 (1.876)	4.791 (8.390)	-1.121 (2.310)	.3551 (2.154)	-1.940 (2.211)	-2.056 (12.77)
		Supplemental White				Native American			
x ₁₉	Not on time in school at age 14	.6094 (.2291)	.2010 (.2162)	-.4738 (.2574)	1.085 (1.099)	.6280 (.3408)	.1195 (.3254)	-.4984 (.3508)	.0765 (1.959)
x ₂₀	Non-urban at age 14	-.1187 (.1826)	-.1088 (.1680)	-.2112 (.2083)	-1.838 (.8520)	.3954 (.2784)	-.2202 (.2886)	.2683 (.3259)	-.9039 (1.721)
x ₂₁	No employed males in household at age 14	-.2960 (.1687)	-.1262 (.1598)	.0458 (.2088)	.3339 (.8280)	.0428 (.2978)	-.2581 (.3102)	.0500 (.3260)	.6362 (1.789)
x ₂₂	Cigarette price index	-1.570 (2.021)	-.5916 (1.937)	-5.331 (2.213)	-3.502 (10.63)	-2.445 (1.870)	.7501 (1.816)	1.892 (2.052)	3.818 (10.95)
x ₂₃	Alcohol price index	-.9506 (2.553)	.7100 (2.591)	.2774 (2.737)	-.8401 (14.89)	1.731 (2.743)	.2797 (2.681)	1.816 (2.798)	-7.113 (16.68)
x ₂₄	Medical services price index	-1.202 (2.631)	1.333 (2.577)	.3052 (2.772)	-11.80 (14.59)	-4.985 (2.652)	-.9403 (2.730)	1.697 (2.750)	-3.191 (16.27)
x ₂₅	Food price index	3.012 (2.467)	2.069 (2.442)	-1.284 (2.564)	8.476 (13.73)	-1.435 (2.764)	1.045 (2.787)	1.016 (2.794)	2.494 (16.39)

	S	D	PC	WG
		Hispanic		
x ₁₉ Not on time in school at age 14	.1251 (.2227)	-.1809 (.1910)	.0259 (.1808)	2.048 (1.013)
x ₂₀ Non-urban at age 14	-.1569 (.2641)	-.3996 (.2143)	-.3059 (.2000)	.9139 (1.113)
x ₂₁ No employed males in household at age 14	-.2190 (.2025)	-.1093 (.1616)	-.0445 (.1699)	-.8069 (.8853)
x ₂₂ Cigarette price index	.6106 (1.546)	.5791 (1.312)	.0583 (1.392)	-4.908 (7.489)
x ₂₃ Alcohol price index	2.656 (2.429)	-.6047 (2.342)	-.3408 (2.368)	.6341 (13.34)
x ₂₄ Medical services price index	1.655 (2.456)	-1.177 (2.276)	.4610 (2.355)	-.9498 (13.39)
x ₂₅ Food price index	3.139 (2.439)	1.139 (2.223)	-.6822 (2.294)	-2.901 (12.95)
		Prior		
x ₁₉ Not on time in school at age 14	.6000 (.4400)	.6000 (.4400)	-.6000 (.4400)	-.6000 (2.640)
x ₂₀ Non-urban at age 14	.0000 (.4400)	.0000 (.4400)	.0000 (.4400)	.0000 (2.640)
x ₂₁ No employed males in household at age 14	.0000 (.4400)	.0000 (.4400)	.0000 (.4400)	.0000 (2.640)
x ₂₂ Cigarette price index	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (7.348)
x ₂₃ Alcohol price index	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (7.348)
x ₂₄ Medical services price index	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (7.348)
x ₂₅ Food price index	.0000 (3.000)	.0000 (3.000)	.0000 (3.000)	.0000 (7.348)

Table 3: Gestation equation by group: posterior means (standard deviations) of Γ_1 and Δ_2 under H_*

	Prior	Main White	Supp. White	Black	Hisp.	Native Amer.
S	-1.000 (2.000)	.4480 (.3880)	-.3833 (.6904)	.2315 (.8059)	-.1121 (.7212)	-.1967 (1.094)
D	.0000 (2.000)	1.255 (.4553)	.7983 (.6243)	-1.880 (.5406)	-.9589 (1.017)	-.7112 (.9714)
PC	.0000 (2.000)	.2368 (.5071)	.4575 (.6929)	1.933 (.9949)	1.138 (.9766)	-.3798 (1.101)
WG	.0000 (2.000)	.0254 (.0255)	.0395 (.0345)	.0193 (.0254)	-.0082 (.0296)	.0061 (.0430)
Intercept	40.00 (1.760)	37.67 (.5938)	37.81 (.8097)	37.43 (.7886)	38.53 (.7468)	39.55 (1.055)
Male child	.0000 (1.760)	-.2455 (.1477)	-.1033 (.3206)	.2075 (.2461)	-.1696 (.2810)	.4540 (.6229)
Mother's age — 23yrs.	.0000 (12.00)	-.1058 (.0213)	-.0425 (.0524)	-.0506 (.0340)	-.1128 (.0444)	-.0178 (.0844)
Body mass index — 24	.0000 (12.00)	.0260 (.2369)	-.6877 (.4426)	.0113 (.2333)	-.0510 (.3977)	-.7709 (.8177)
Maternal height — 162cm	.0000 (12.00)	-.0013 (.0685)	-.1950 (.1264)	.0166 (.0720)	.0087 (.1194)	-.3070 (.2579)
Maternal weight — 63kg	.0000 (12.00)	.0068 (.0873)	.2514 (.1636)	.0197 (.0861)	.0406 (.1543)	.2946 (.3075)
Height	-.0430 (12.44)	-.0090 (.0135)	.0088 (.0266)	.0133 (.0185)	.0239 (.0259)	-.0786 (.0556)
Weight	.1599 (13.03)	.0167 (.0074)	-.0106 (.0164)	.0240 (.0103)	.0212 (.0154)	.0008 (.0264)

Table 4: Birth length equation by group: posterior means (standard deviations) of Γ_1 , Γ_2 and Δ_2 under H^*

	Prior	Main White	Supp. White	Black	Hisp.	Native Amer.
S	.0000 (3.000)	-1.581 (.5443)	-1.810 (.9062)	-.4231 (1.541)	-1.055 (1.435)	-.1968 (1.086)
D	.0000 (3.000)	-1.283 (.5805)	-1.385 (1.100)	1.341 (1.106)	1.201 (1.398)	-.8107 (1.202)
PC	.0000 (3.000)	1.442 (1.011)	-1.828 (1.017)	.2599 (1.077)	-1.129 (1.235)	.1126 (1.319)
WG	.1000 (1.000)	.0605 (.0342)	.0980 (.0431)	.1325 (.0412)	.0336 (.0465)	.1026 (.0456)
G	.0500 (1.000)	.0884 (.0527)	.1338 (.0532)	.0069 (.0529)	.0936 (.0545)	.0726 (.0555)
Intercept	48.00 (1.760)	46.69 (1.787)	47.11 (1.714)	47.39 (1.749)	46.79 (1.756)	47.35 (1.705)
Male child	.1000 (1.760)	.7303 (.2283)	.9287 (.4716)	.8529 (.4873)	1.120 (.5107)	.4028 (.6407)
Mother's age — 23yrs.	.0000 (12.00)	-.0933 (.0326)	.0770 (.0772)	.0991 (.0664)	-.0440 (.0816)	-.0429 (.0893)
Body mass index — 24	.0000 (12.00)	.0535 (.0277)	-.0024 (.0620)	.0351 (.0527)	.0441 (.0704)	.0625 (.0706)
Maternal height — 162cm	.0000 (12.00)	.0854 (.0185)	.0941 (.0355)	.0455 (.0353)	.1294 (.0448)	.0387 (.0532)
Height	-.4611 (12.22)	.0696 (.0196)	.0948 (.0388)	.0351 (.0362)	.1163 (.0468)	.0202 (.0569)
Weight	.0928 (4.543)	.0204 (.0105)	-.0009 (.0236)	.0134 (.0201)	.0168 (.0268)	.0238 (.0269)

Table 7: Birth weight equation by group: posterior means (standard deviations) of τ_1 , τ_2 and Δ_2 under H_*

	Prior	Main White	Supp. White	Black	Hisp.	Native Amer.
S	-.3500 (1.000)	-.2887 (.1008)	-.6074 (.1859)	-.0482 (.2442)	-.1299 (.2045)	-.1136 (.2908)
D	.0000 (1.000)	.1239 (.1418)	.1536 (.1435)	-.1448 (.1684)	-.0763 (.2531)	.0268 (.3135)
PC	.1000 (1.000)	.1192 (.1443)	-.0424 (.1906)	.2244 (.2074)	.3977 (.2448)	-.0015 (.3246)
WG	.1000 (1.000)	.0009 (.0155)	-.0222 (.0137)	.0182 (.0125)	.0070 (.0133)	.0036 (.0167)
G	.0100 (1.000)	.0363 (.0210)	.0471 (.0241)	.0226 (.0244)	.0296 (.0224)	.0362 (.0231)
Intercept	2.000 (.8800)	1.820 (.7927)	1.843 (.8940)	1.930 (.9033)	1.806 (.8472)	1.857 (.8496)
Male child	.1000 (.8800)	.0920 (.0351)	.2065 (.0758)	.1117 (.0579)	-.0124 (.0657)	.0765 (.1531)
Mother's age — 23yrs.	.0000 (6.000)	-.0186 (.0055)	-.0194 (.0128)	.0014 (.0079)	-.0137 (.0108)	-.0066 (.0209)
Body mass index — 24	.0000 (6.000)	-.0353 (.0117)	-.0192 (.0208)	-.0126 (.0139)	-.0477 (.0208)	-.0005 (.0409)
Maternal weight — 63kg	.0000 (6.000)	.0207 (.0044)	.0135 (.0071)	.0083 (.0052)	.0241 (.0080)	.0191 (.0144)
Height	.0003 (1.797)	.0105 (.0035)	.0057 (.0062)	.0037 (.0041)	.0141 (.0062)	.0001 (.0121)
Weight	-.0923 (6.443)	.0072 (.0017)	.0062 (.0037)	.0035 (.0025)	.0059 (.0036)	.0189 (.0062)

Figure 1: Posterior predictive distribution for gestation under H_*

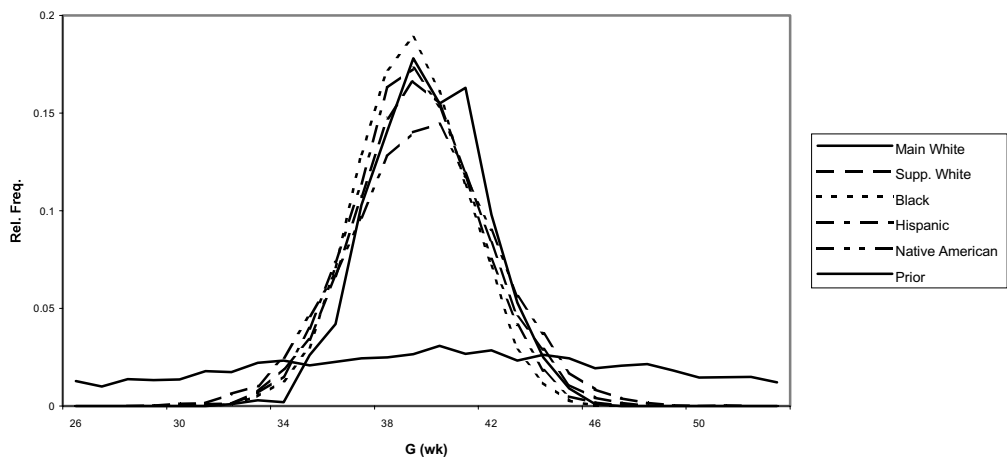


Figure 2: Posterior predictive distribution for birth length under H_*

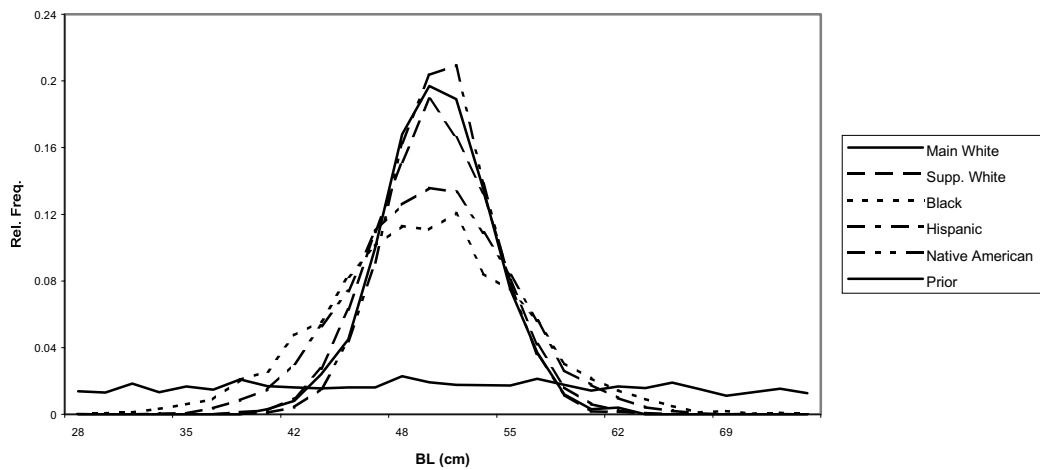
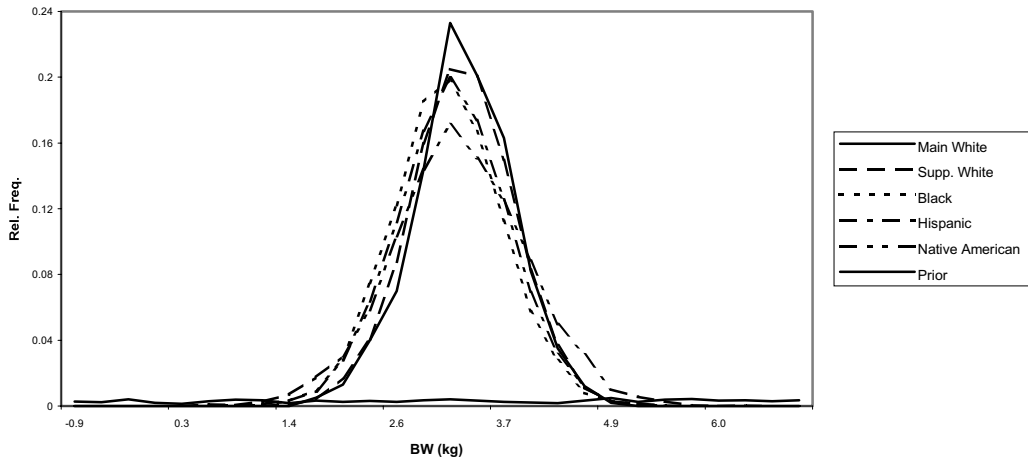


Figure 3: Posterior predictive distribution for birth weight under H_* 

7. References

- [1] Chib, S. and Greenberg, E. (1998), 'Analysis of multivariate probit models', *Biometrika*, Vol. 85, pp. 347–361.
- [2] Grossman, M. (1972), 'On the concept of health capital and the demand for health', *Journal of Political Economy*, 80, pp. 223–255.
- [3] Illsley, R. and Mitchell, R. G. (1984), 'The developing concept of low birth weight and the present state of knowledge', in Illsley, R. and Mitchell, R. G. (eds), *Low birth weight: A medical, psychological and social study*, Wiley, New York, pp. 5–32.
- [4] Institute of Medicine (1985), *Preventing low birth weight*, Committee to Study the Prevention of Low Birth Weight, Division of Health Promotion and Disease Prevention, National Academy Press, Washington, DC.
- [5] Li, K. (1998), 'Bayesian inference in a simultaneous equation model with limited dependent variables', *Journal of Econometrics*, 85, pp. 387–400.
- [6] Li, K. and Poirier, D. J. (2001), 'An econometric analysis of the birth process by racial/ethnic groups', in George, E. (ed.), *Bayesian Methods with Applications to Science, Policy, and Official Statistics (Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis)*, pp. 21–30.
- [7] Li, K. and Poirier, D. J. (2000), 'An econometric model of birth inputs and outputs: A detailed report', unpublished manuscript, University of British Columbia, available at <http://finance.commerce.ubc.ca/research/abstracts/UBCFIN00-3.html>.

- [8] Li, K. and Poirier D. J. (1999), 'An econometric model of birth weight for native Americans', unpublished manuscript, University of British Columbia, available at <http://finance.commerce.ubc.ca/research/abstracts/UBCFIN99-7.html>.
- [9] Miller, H. C. and Merritt, T. A. (1979), *Foetal growth in humans*, Year Book Medical Publishers, Chicago.
- [10] Poirier, D. J. (1995), *Intermediate statistics and econometrics: A comparative approach*, MIT Press, Cambridge.
- [11] Poirier, D. J. (1998), 'Modelling birth weight: US stylised facts and a window for viewing them', unpublished manuscript, University of Toronto.

Bayesian estimation in a US Census Bureau survey of income recall using respondent-generated intervals

S James Press (*) and Kent H. Marquis (**)

(*) University of California, Riverside, CA 92521-0138 (jpress@ucr.ac1.ucr.edu)

(**) US Bureau of Census, Washington, DC 0233-9000

(kent_hammond_marquis@ccmail.census.gov)

Keywords: Bayesian, estimation, surveys, respondent, intervals, Census Bureau, bias, non-sampling errors, recall, income

Abstract

This paper presents some preliminary results of a US Census Bureau research telephone survey of income in a national household sample. The survey questionnaire used a survey procedure called respondent-generated intervals (RGIs) which requests that respondents supply bounds on their estimates of the sizes of recalled quantities. A three-stage Bayesian hierarchical model, as well as several alternative estimators, were used to estimate population means for selected income-related items. The bounds information was used to assess the hyperparameters of the prior distributions. Questionnaire design and testing was carried out as a joint effort of statisticians and cognitive scientists using the cognitive laboratory facilities of the US Census Bureau. A Markov Chain Monte Carlo, Gibbs sampler was used to find Bayesian estimates of population means. The sample median and Bayesian estimators proved to be the best estimators relative to the benchmark we used for establishing true values (income tax forms). An interval estimator based upon the bounds information was the best of several considered, and covered the true value two thirds of the time.

1. Introduction

This paper presents some results of a sample survey that was carried out by the US Census Bureau to explore the efficacy of the “Respondent-Generated Intervals” (RGI) survey protocol when used with sensitive questions. The idea is to use upper and lower bounds information to help to estimate population means in questions of recall of factual information. We explored questions involving recall of various types of household income.

The basic problem is how to improve population estimates from surveys or censuses when the responses contain bias errors, and the distribution characteristics of the response errors are unknown. It seemed to us that bias arises from individual differences, and Bayesian methods can be used to explore individual differences. The notion is to elicit information from respondents about their recall distributions for some fact, just as we might elicit information about their prior distributions for some unknown quantity. We use the information about their recall distributions to develop an empirical Bayes estimator of the population mean. Perhaps such an estimator is more accurate, has less variability than a traditional estimator of the population mean, and is accompanied by an increase in response rate. The RGI protocol reminds the respondent of the fallibility of memory and then requests that the respondent to a recall question give not only a basic response (called

the ‘usage’ quantity), but also, a lower bound and an upper bound for the smallest and largest values the true answer to the recall question might possibly be for that respondent. This auxiliary bounds-information is then used to assess hyperparameters of the prior distributions associated with an empirical Bayesian estimator of the population mean. To study the properties of such an estimator it was decided to use the RGI protocol in a telephone survey of income-related questions that are normally considered ‘sensitive’, in that respondents may tend to misrepresent their responses, in order to conceal their true incomes. We could not use simulation of any sort to study the properties of such an approach because results depend so heavily on the real behaviour of human subjects. Therefore, to determine the accuracy of the new estimator, we compared the estimated population means with the ‘true values,’ obtained from income tax forms.

The plan of the paper is to present the background and origins of the current research in Section 2. In Section 3 we discuss the design of the research survey we conducted including the design of the questionnaires, the cognitive testing of the questions, the survey instrument format, the telephone survey, and the problems we had getting respondents to comprehend the questions. Section 4 is devoted to explaining the data cleaning procedures we used. Section 5 presents some cognitive/statistical modelling of the data and the modelling of the Bayesian estimator. We conclude with the results and some conclusions in Section 6.

2. Related work

The RGI protocol for questionnaire design has its origins in Bayesian assessment procedures wherein an entire prior distribution (and/or a utility function) for an individual is assessed by connecting a collection of points on the individual’s distribution by means of a sequence of elicitation questions (see, for example, Schlaifer, 1959, Chapter 6; and Hogarth, 1980, Appendices B and C). Proceeding along related lines, the RGI protocol was proposed, and a Bayesian estimation model was developed (see Press, 1999). It was later used in experiments embedded in two different, but related, sample surveys carried out on two university campuses. One was carried out on the University of California, Riverside, campus, and the other on the State University of New York, Stony Brook, campus (for details, see Press and Tanur, 2000a). Undergraduate college students on the Riverside campus were given eight recall items and on the Stony Brook campus they were given 10 recall items, for a total of 18 items. Respondents were asked to recall items relating to student affairs, such as grade point averages. True values on the two campuses were verified by the university administrations. Results suggested there were gains in estimation accuracy that could be achieved by the new procedure. The Bayesian estimate was the most, or second most, accurate estimator of four estimators being compared over 18 items. There were also suggested gains in the response rate (see Press and Tanur, 2000b). But there were still many questions remaining regarding issues such as how well would such a protocol work with sensitive questions, such as ‘income’? For related work, see Kennickel (1997), who described the 1995 Survey of Consumer Finances (SCF), carried out by the National Opinion Research Centre in Chicago, as including opportunities for the respondents who answered either ‘don’t know’, or ‘refuse’, to select

from eight pre-assigned ranges, or to provide their own upper and lower bounds ('volunteered ranges'); and the 'unfolding brackets' approach used in the Health and Retirement Survey, see Heeringa et al. (1995); and Juster, F.T. and J.P. Smith (1996). For details about the cognitive aspects of this work, see Marquis and Press (1999). See also, Schwartz and Paulin (2000) for a comparison of these methods.

3. Design of the experiment

Our initial goal was to develop questioning procedures to elicit the standard answer and the range of plausible alternative values. For estimation purposes, we wanted to get quantitative, interval scale information useful in fitting a Bayesian prior distribution for each respondent. So we decided to ask about income. To cover a range of difficulties, we asked about two types of income for the most recent calendar year (1997) and the year before that. Then we asked how much each of the two types of income had changed over the past five years (a very difficult cognitive task). The income types were wages and salaries on the one hand and interest and dividends on the other.

The goal of the telephone survey was to obtain a best estimate report of an income amount and a report of the uncertainty range surrounding the estimated amount for several income items. These data are used in Section 5 to develop improved estimation procedures.

Sample

We developed a frame of households from the Census Bureau's commercial and administrative records containing households that filed joint tax returns having wage and salary income for the last five consecutive years. The frame covered the four states in which the American Community Survey (ACS) held its first pilot tests. Households interviewed in the ACS tests or for which we could not obtain current phone numbers were eliminated from the frame. A sample of about 2 000 households was drawn from this frame, and each was assigned to an experimental interviewing treatment. The Census Bureau obtained a quota sample of 500 completed interviews, eliminating households that had become ineligible through retirement, death, divorce or other circumstances that precluded observing the joint wage and salary income on the tax return.

We used two versions of the questionnaire. Each version asked about wage and salary income and about interest and dividend income for three time periods: the calendar years 1997, and 1996, and the amount of income changes over the last five years (1993–97). Both versions included questions about characteristics that might correlate with income reporting accuracy, such as: Who pays the bills? Who fills out the federal tax form?, level of education and age.

Version one of the questionnaire, administered to 75 % of the eligible, completed cases, asked for the low-range boundary first, then the high-range boundary, then the best estimate. Version two, administered to 25 % of the eligible, completed cases, reversed the

ordering by asking first for the best estimate of the income amount, then the confidence rating, then the lower-bound estimate, and finally, the upper-bound estimate.

Telephone interviewing was conducted in May and June of 1998. We held a half-day training session for the telephone interviewers, covering the procedures and concepts, and providing detailed income definition information in case respondents asked about special circumstances.

Since the frame information also included data from administrative records about household income, we eventually linked the survey responses to the administrative records to evaluate the validity of the telephone survey responses.

We had about 2 000 potential respondents and actually obtained about 500 completed questionnaires provided by the telephone interviewers who used Computer Assisted Telephone Interviewing (CATI) computer software. There were six basic income-related questions in each of the two versions of the questionnaire for a total of 12 basic items. For each of the 12 items we had asked for three responses: a basic, traditional response ('best estimate'); a lower bound; and an upper bound. In addition, we asked some demographic types of questions:

Who in your household usually handles the household finances like paying the bills?

Who is the person who usually does the federal income taxes?

How old were you on your last birthday?

How old was your spouse on his/her last birthday?

What was the highest level of schooling you completed, or the highest degree you have received?

Did you earn income in 1997? Did your spouse earn income in 1997?

Would you say that you and your family are better off or worse off financially than you were a year ago?

In version two of the questionnaires we also asked the 'confidence' question:

How sure are you that your estimate is the correct actual value? Would you say it is probably very close, probably very far away, or probably somewhere between close and far away?

Before any data analysis could take place the data had to be 'cleaned', that is, certain respondents and certain responses had to be filtered out first. There were three issues and steps involved in the data cleaning process.

4. Modelling and analysis

Using the data from this survey, for each item, we compared five point estimators:

- the sample mean,
- the midpoint estimator,
- the sample median estimator,

- the Bayesian estimator whose prior mean is taken to be the sample median,
- the Bayesian estimator whose prior mean is taken to be the midpoint estimator.

We also developed three interval estimators. These estimators are explained below.

The basic data is really a triple of observations for each respondent: a usage quantity response, and upper and lower bounds. The triples of data from the respondents are considered to be mutually independent across respondents. But because each respondent is attempting to recall a different income parameter, the data triples cannot be identically distributed. A classical, or sampling theory approach to modelling the usage and bounds data for a given respondent does not seem feasible since it is not clear how to model the joint distribution of the data. Alternatively, for a given respondent, we'll treat the usage quantity as basic data, and then use the bounds information as auxiliary, to assess the hyperparameters of prior distributions in a Bayesian hierarchical model.

4.1. Mid-point estimator

For any given item, let a_i and b_i denote the lower and upper bounds given by respondent i , respectively, and let X_i denote respondent i 's usage quantity. The mid point of the interval (a_i, b_i) is given by $(a_i + b_i)/2$. This is, we believe, an intuitively-sensible estimator of the income parameter that respondent i is attempting to recall. If we average these mid points with equal weights, we call the result the mid-point estimator of the population mean. We also note that some respondents gave symmetric responses while others gave asymmetric responses. That is, if the usage quantity were in the middle of the interval given we called the response symmetric; otherwise, asymmetric. For symmetric responses, the usage quantity X_i was in the middle of the interval. The sample mean and the mid-point estimator would be identical therefore if all responses were symmetric; but they weren't. In fact, while percentages varied across items, typically, for a given item, most of the responses were indeed symmetric: the percentages of mid-point (symmetric) responses exceeded the percentages of asymmetric responses in eight out of the 12 items (Items 1,5,7-12). In five of the 12 items (Items 5,8,10,11,12), the percentages of asymmetric responses tailed to the left more often than to the right. In seven of the 12 items (Items 1, 2, 3, 4, 6,7,9), the percentages of asymmetric responses tailed to the right more often than to the left. (see Table 1). We average the end points to form \bar{a} , and \bar{b} , for the average lower and upper bounds, respectively, and then form the interval (\bar{a}, \bar{b}) . We refer to this interval as the Average Respondent Generated Interval, or ARG I. The evidence suggests that the ARG I is a reasonable interval estimator for the population mean, and is a strong competitor of confidence or Bayesian credibility intervals. We'll compare the numerical values of these interval estimators in Section 5.

4.2. Bayesian estimators

In a typical Bayesian analysis involving an issue for which certain people have some special knowledge, we often assume before taking any data that the subject has a (prior)

distribution about the unknown quantity in his/her head, and it is the task of the analyst to elicit some points on that prior distribution. The elicited prior distribution is then combined with the likelihood for the data via Bayes' theorem to develop the posterior distribution from which inferences about the unknown quantity will be made. Analogously, we will assume in this problem that each respondent to a recall question really has a complete 'recall distribution' in his/her head for the answer, but the respondent is requested by the interviewer to give just three points on that distribution, the respondent's best guess about the location of the distribution (the respondent's recalled usage quantity) and one extreme value in each of the tails (the respondent's lower and upper bounds). The distribution may be symmetric, or not. We now develop a three-stage Bayesian hierarchical model for the population mean. Our approach will be to incorporate the usage quantities into the likelihood function for the data, and to utilise the bounds information to assess the parameters of the various prior distributions (hyperparameters), and the data-distribution variances. Because we are using the bounds information to assess the hyperparameters, the resulting estimator is really an empirical Bayes' estimator.

The Normal-Normal hierarchical model

We start with the model for the usage quantity for the recall distributions. We are concerned here not with the distribution of the data taken as if we had independent, identically distributed data from a common distribution. Instead, we are concerned with modelling the distribution of each respondent's recall distribution. As mentioned earlier, we have only three points on each distribution, so we can hardly fit a model. But as mentioned earlier (Section 5.1), we have examined the degree of symmetry in each distribution (see Table 1) and have found that a normal distribution assumption is not inconsistent with these data. Accordingly, assume that conditionally, the X_i 's, the usage quantities, are mutually independent, and $(X_i | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2)$. Define the $(n \times 1)$ column vectors:

$X = (X_i)$, $x = (x_i)$, $\theta = (\theta_i)$. We are able to assume the σ_i 's are approximately known by using the bounds information. We estimate them by assuming that for a normal distribution, almost the total mass of the distribution is included within three standard deviations of the mean. We therefore estimate them by taking: $\sigma_i = (b_i - a_i) / 6$, where again, a_i and b_i denote the lower and upper bounds supplied by respondent i .

For the joint distribution of the means of the recall distributions, we take the multivariate normal distribution whose means are located at the common mean θ_0 (for this group of respondents), with common variance τ^2 . That is,

$$p_\theta(\theta / \theta_0, \tau^2) = \frac{1}{(2\pi)^{n/2}} \exp \left\{ -\frac{1}{2\tau^2} (\theta - \theta_0 I_n)' (\theta - \theta_0 I_n) \right\}$$

where I_n denotes the identity matrix of order n . Only θ_0 and τ^2 remain to be modelled. Let $\bar{\theta}_0$ denote the mean of the θ_0 distribution (the prior mean), and assume:

$$(\theta_0 | \bar{\theta}_0, \delta^2) \sim N(\bar{\theta}_0, \delta^2), \quad \text{and} \quad p(h | \phi) = \phi \exp\{-\phi h\},$$

where: $h \equiv \tau^{-2}$ denotes the precision of the distribution of θ_i . By Bayes' theorem, the posterior probability density function of θ_0 , given x , is given by:

$$p(\theta_0 | x) = \frac{\iint p(x | \theta) p(\theta | \theta_0, h) p(\theta_0 | \bar{\theta}_0, \delta^2) p(h | \bar{\theta}_0, \delta^2) d\theta dh}{\iiint p(x | \theta) p(\theta | \theta_0, h) p(\theta_0 | \bar{\theta}_0, \delta^2) p(h | \bar{\theta}_0, \delta^2) d\theta dh d\theta_0}.$$

The Bayes' estimator of the population mean is taken to be the posterior mean:

$$\bar{\theta}_0^* = E[\theta_0 | x] = \int \theta_0 p(\theta_0 | x) d\theta_0.$$

The numerical evaluation of the Bayes estimator is effected by means of the Gibbs sampler. The one-dimensional conditional densities of all of the variables are readily obtained by conditioning in the joint density. This is all we need to be able to apply the Gibbs sampler since there are no improper distributions being used, and because we know that the joint density exists (we know it explicitly). To run the Gibbs sampler on our data we used the WinBUGS program Version 1.2 (see Spiegelhalter et al., 1999). Conveniently, it was not necessary to work out the joint or conditional distributions in any particular format, but only necessary to input the three stages of distributions using the proper WinBugs syntax.

5. Results and conclusions

5.1. Results regarding accuracy

We define accuracy in terms of how close the survey response is to the entries on the family's federal income tax form (such information is available to the Census Bureau on a very limited and very restricted confidential basis for research purposes only). While we recognise that it is sometimes claimed that there are those who underestimate their incomes for tax purposes, we nevertheless have taken the income tax statement as our gold standard for 'true' household incomes.

Many results of this analysis are given in Table 2a, which is devoted to a comparison of point estimates, and in Table 2b, which is devoted to a comparison of interval estimates.

The first column of Table 2a denotes the case number of the 12 cases corresponding to the 12 income-related items in the questionnaires. The odd-numbered case numbers refer to questions in the version one questionnaire; the even-numbered case numbers refer to the corresponding questions in the version two questionnaire. The odd-numbered questions differ from the even-numbered questions mainly in the ordering of the subsidiary questions. For example, in case 1, the bounds questions were asked first (the lower bound was always requested first, and then the upper bound), and then the basic usage question was asked. In case 2, the order was reversed, so that first the same basic question (usage) was asked, but then the two bounds questions were asked. Analogously for cases 3 and 4; and then cases 5 and 6; etc.

Under sample size (column 3) there were two numbers indicated for each case. The first is the sample size used to compute the Bayes estimates, and the second is the sample size used to compute all other estimates. The reason they differ is that there were generally instances for which the respondent gave the same response for an upper bound, and for a usage quantity, and for the lower bound. In such instances where the respondent did not indicate any uncertainty about his/her recall, the estimated variance in the data for that respondent was zero. It was therefore inappropriate to model a Bayesian estimator with a normal likelihood for a data point with zero variance. Such responses were still used to calculate the sample mean, the sample median, and the mid-point estimator, but these responses were not included in the Bayesian estimator calculations.

Column 4 of Table 2a gives the true value of the population mean (the mean value of all the values presented on the tax returns for all respondents in our survey, for each item). Column 5 in Table 2a gives the usual sample mean estimator, in dollars (the average usage quantities), and column 6 gives the Gibbs sampler-computed Bayes estimates of the population mean, first using the sample median as the prior mean, $\bar{\theta}_0$, and then using the mid-point estimator as the prior mean, $\bar{\theta}_0$. The Gibbs sampler was generally run for 100 000 iterations, or until it was clear that convergence had taken place. The posterior

probability density function for θ_0 was always unimodal and well-behaved with convergence taking place rapidly.

Column 7 gives the mid-point estimator. The last two columns of Table 2a list the variances of the lower and upper bounds, averaged across respondents.

In Table 2b we repeat the first four columns of Table 2a, for convenience. Then, the next three columns give the three interval estimates being compared: the ARG1, the interval ranging from the mean lower bound to the mean upper bound, the 95 % confidence interval, and the 95 % Bayesian credibility interval. Finally, the last three columns give the lengths of these three intervals.

Next we examine the ordering effect of whether it matters if the usage question is asked first, or after the bounds questions. We may see from Table 2b that the ARG1 is always smaller when the questions are ordered so that the basic usage question is asked first, followed by the bounds questions, for cases 1-6. These first six are the cases where the information is probably best known (perhaps respondents give shorter intervals when they can utilise their usage response as an anchor, if they are confident of their usage response). The trend is exactly reversed for the last six cases where the information is probably not as well known. In these last six cases, the shorter ARG1 is found by asking the bounds questions before the basic usage question.

We also note that in both tables, the sample sizes differ from the odd-numbered cases to the even-numbered cases by a ratio of approximately 3:1 from odd number to even number (from version one to version two). These sample size differences will certainly affect the sizes of the ARG1's, but they shouldn't have very much of an effect because all of the sample sizes (the ones in parenthesis in Table 2a) are at least about 90, as compared with the larger sample sizes of about 300. Consistency of the lower and upper bound means in this roughly symmetric situation ensures that the sample size differences will have but a minor effect.

In Table 2b we see from the last three columns that the 95 % Bayesian credibility interval is always shorter than both the 95 % confidence interval, and the ARG1. Moreover, the ARG1 and the 95 % confidence intervals are each shorter than one another for about half of the items. We may also note that both the 95 % Bayesian credibility interval and the ARG1 are shorter for the usage question being asked first, and the bounds questions following, for the 1997 salary and wages, and interest and dividends (the most recent information); the effect reverses for the less recent information. (In cases 5 and 6 there is a small inconsistency of this trend between the credibility interval the ARG1, but the difference is small.)

Examining the lengths of the 95 % confidence intervals, asking the basic usage question first, followed by the bounds questions, always resulted in longer confidence intervals than were found by asking the questions in the reverse order, in disagreement with both the Bayesian credibility intervals and the ARG1's. At least part of the differences in length of

the confidence intervals has to be attributable to relative sample sizes; recall that the version that asked for bounds first followed by usage typically had sample sizes three times that of the version that asked for the standard question followed by the bounds. Moreover, the 95 % confidence interval does not depend upon the bounds information in any way, whereas both the Bayesian credibility interval and the ARG1 do take the bounds into account; so it's not surprising that the ordering of the questions would reflect this difference. The 95 % confidence interval is always smaller for larger sample sizes (an effect that is confounded with the ordering effect).

We were disappointed to find that for these very sensitive income questions, none of our competing point estimators performed very well with respect to accuracy when the item requested recall of the size of a change in some type of income over a five-year period (cases 5, 6, 11, 12). There were five point estimators: the sample mean, the sample median, the two Bayesian estimators, and the mid-point estimator. Each was the most accurate of the five for a few times out of the 12 cases, and none was the most accurate more than five times out of the 12 cases. For 1996 and 1997 salary and wages (cases 1-4), the true (tax form) value was always less than any of the estimators. For 1996 and 1997 interest and dividends (cases 7-10), the point estimators did not always overestimate the true values.

Things looked somewhat better for interval estimation, however. There were three interval estimators being compared: the 95 % confidence interval, the 95 % Bayesian credibility interval, and the ARG1 (95 % was selected for comparison purposes only because it is the most often-used level). The ARG1, or average respondent-generated interval (interval from the average lower bound to the average upper bound) covered the true values eight times in the 12 cases, and the 95 % confidence interval covered the true values seven times in the 12 cases. The only times the ARG1 did not cover the true values was for the two, five-year change items: How much has your household income from salary and wages changed over the last five years? How much has your household income from interest and dividends changed over the last five years? It didn't seem to matter in which order the bounds questions were asked. Regardless of order, for the four questions covering change in income over the last five years, all of the estimators had difficulty covering the true value. These were the items that were the most difficult to recall since they required more complex cognitive tasks. Depending upon the strategy the respondent used to answer such a question, and the interpretation the respondent gave to the meaning of the question, it may have been necessary for a respondent to recall not only the income value today, and the income value five years ago, but then to calculate the difference. Comparing the last three columns in Table 2b shows that the Bayesian 95 % credibility intervals were always shorter than the other two intervals. This was of course not surprising since such shorter intervals generally result when credibility intervals are calculated using non-vague ('informative') prior distributions.

Averaged over the 12 items, the root mean squared percentage error was smallest for the median usage estimator, followed by both Bayesian estimators (see Table 3). It is likely that improved assessment of the hyperparameters in the hierarchical model will improve the performance of the Bayesian estimator. It is also likely that had

hyperparameters been assessed from results of earlier surveys, numerical results would have been better.

5.2. Results regarding non-response

We also examined response rate. There were generally many instances in which respondents did not supply usage quantities, but mostly, in such cases, they also did not supply bounds information. In fact, the number of instances in which respondents provided (both) bounds information when they did not supply usage quantities ranged from 0 to 4, across the 12 items we studied. This result is in stark contrast to results obtained in earlier experiments where percentages of instances for which respondents gave bounds information when they did not supply usage quantities ranged as high as 41 %. These earlier experiments (on university campuses) did not involve such sensitive (income-related) questions. Moreover, they involved pencil and paper questionnaires instead of the telephone survey we are discussing here, and they involved undergraduate student respondents rather than respondents from established households, who have been presented with questions from professional interviewers representing the US Census Bureau. Overall, there was greater respondent cooperation in this government survey by telephone than we found in our earlier campus-based experiments. We are not certain whether the cooperation effect is attributable to the fact that this was a US Census Bureau survey, or whether it is attributable to the nature of the administration of the survey.

5.3. Conclusions

We used income tax returns as the gold standard for truth. This approach presents several problems. First of all, unlike the usual record check surveys that use administrative record checks as the gold standard, our procedure compares the official self-reports in the tax form with the self-reports in the survey. These two situations of self-reporting may well generate different pressures for over- and under-reporting. Under-reporting in the tax form results in lower tax assessment but is constrained by respondent's awareness that the Internal Revenue Service receives independent reports from employers and payers of dividend and interest. There are no such constraints on under-reporting in the survey context. Further, it is not in the taxpayer's interest to exaggerate his/her income on a tax form; on the other hand, issues of social desirability may encourage such over-reporting in a survey context. We actually found that the 'true' income tax records provided values that were smaller than any of the point estimates in all but three cases: cases 8, 9, 10.

Secondly, there are possibilities of misinterpretation of what the respondent was to report on the survey. For example, we asked 'What is your best estimate of your combined wages and salary income for 1997?' We were quite clear about the line on the income tax form that we considered 'truth' in this case. But would the respondent interpret the question in the same way? Would he/she report gross income, or take-home pay? Would he/she exclude non-taxable income such as deductions for 'supplemental retirement annuities', or before-tax health insurance premiums?

Thirdly, especially if the respondent uses a professional to prepare his/her taxes, as most of our respondents did, he/she may have very poor knowledge of some of the amounts we asked for; in particular, the totality of interest and dividend income. Dividends that are reinvested automatically, and interest that is automatically credited to bank accounts are usually summarised together with other interest and dividends for respondents only once a year on their tax forms. Unless a respondent carefully studies those forms before signing them, he/she may have little knowledge of the total amount of interest and dividends accumulated.

Finally, our question about change in salaries and wages over a five-year period requires the respondent to recall two amounts each plagued with all of the difficulties detailed above, and then to perform an arithmetic operation. Not only is this task not pure recall, but it is exceedingly daunting cognitively. And the task is even more daunting when we ask for change in interest and dividend income.

For all of these reasons we are not terribly surprised that none of the point estimates came very close to truth as defined by our gold standard.

The RGI protocol provided the ARG1 interval estimates that did cover the ‘true values’ two thirds of the time, in spite of possible systematic under-reporting biases. These fairly successful interval estimates confirm the earlier successful results from the campus experiments. Moreover, they suggest that the RGI protocol is likely to provide improved interval estimates over traditional interval estimates even for sensitive questions.

6. Acknowledgements

The authors are grateful for the technical assistance of Dr Robert Creecy, Rebecca Giem, Meredith Lee, Lily Liang, MaryAnn Scaggs, Dr Judith Tanur and George Woodworth; for the financial support of the US Census Bureau through its ASA/NSF/Census Fellow Programme and through its Research Grant No 43-YA-BC-910579-02; and for the financial support of the University of California, Riverside.

7. Tables

TABLE 1
RESULTS ON SYMMETRY/ASYMMETRY IN CENSUS SURVEY DATA

Case #	Question / order of bounds	Coded response (*)	Count	Per cent	Sample size
1	1997 — salary and wages/ low–high usage	-1	79	22.83	346
		0	120	34.68	
		1	147	42.49	
2	1997 — salary and wages/ low–high usage	-1	38	33.33	114
		0	33	28.95	
		1	43	37.72	
3	1996 — salary and wages/ low–high usage	-1	79	23.03	343
		0	112	32.65	
		1	152	44.31	
4	1996 — salary and wages/ low–high usage	-1	29	25.44	114
		0	33	28.95	
		1	52	45.61	
5	Five-year change in salary and wages/ low–high usage	-1	105	34.88	301
		0	120	39.87	
		1	76	25.25	
6	Five-year change in salary and wages/ low–high usage	-1	30	30.61	98
		0	33	33.67	
		1	35	35.71	
7	1997-interest and dividends/ low–high usage	-1	70	23.65	296
		0	150	50.68	
		1	76	25.68	
8	1997 — interest and dividends/ low– high usage	-1	29	28.71	101
		0	44	43.56	
		1	28	27.72	
9	1996 — interest and dividends/ low– high usage	-1	73	25.26	289
		0	139	48.10	
		1	77	26.64	
10	1996 — interest and dividends/ low– high usage	-1	33	33.33	99
		0	40	40.40	
		1	26	26.26	
11	Five-year change in interest and dividends/ low–high usage	-1	87	31.75	274
		0	140	51.09	
		1	47	17.15	
12	Five-year change in interest and dividends/ low–high usage	-1	35	38.46	91
		0	36	39.56	
		1	20	21.98	

(*)Coded as -1 means usage response is closer to lower bound; coded as +1 means usage response is closer to upper bound; coded as zero means usage response is at the mid point of the range.

TABLE 2a
COMPARISON OF POINT ESTIMATES FOR US CENSUS SURVEY DATA

Case #	Question/ order of bounds	Sample size Bayes est. (sam est)	True value of pop. mean \$	Sample mean est. of pop. mean \$	Bayes' est. with median/ with mid point.	Mid-point estimate of pop. mean \$	Variance of lower bound	Variance of upper bound
1	1997— salary and wages/ low- high usage	300 (346)	51,586	58,377.0 (*)	51,350 (*) 59,050	58,725.4	2,010,893,894	5,930,461,659
2	1997— salary and wages/ low- high usage	99 (114)	47,766	51,776.7	50,360 50,670	50,580.2	1,045,668,181	1,301,852,072
3	1996— salary and wages/ low- high usage	318 (343)	48,927	55,294.5	56,690 56,020	53,758.7 (*)	1,184,415,986	2,304,317,481
4	1996— salary and wages/ low- high usage	104 (114)	44,872	50,213.6	47,090 49,150	49,120.4	1,026,688,959	1,337,400,169
5	Five-year change in salary and wages/ low- high usage	275 (301)	7,339	17,438.4	17,420 (*) 17,410	18,267.2	269,767,209	760,673,311
6	Five-year change in salary and wages/ low- high usage	93 (98)	3,088	17,244.9	14,220 14,220	16,486.3	337,607,773	759,759,802
7	1997— interest and dividends/ low-high usage	231 (292)	2,589	2742.8 (*)	409.8 2853	2853.7	65,993,805	129,611,824
8	1997— interest and dividends/ low-high usage	84 (101)	3,881	3803.4 (*)	406.1 3985	3985.2	48,725,782	135,833,378
9	1996— interest and dividends/ low-high usage	234 (289)	2,346	2081.3	307.4 2139	2140.2	29,930,325	80,703,238
10	1996— interest and dividends / low-high usage	88 (99)	3,901	3621.8	306.3 3750	3750.3 (*)	38,898,562	108,607,757
11	Five-year change in interest and dividends/ low-high usage	219 (274)	912	1574.8 (*)	154.4 1804	1804.0	14,894,850	72,111,462

12	Five-year change in interest and dividends/ low-high usage	76 (91)	1,124	3543.1 (*)	214.1 (*) 4160	4174.5	71,587,123	990,988,338
----	--	---------	-------	------------	-------------------	--------	------------	-------------

(*) Most accurate among the four point estimators.

TABLE 2b
COMPARISON OF INTERVAL ESTIMATES FOR US CENSUS SURVEY DATA

Case #	Question/ order of bounds	Sample size Bayes' est. (sam est)	True value of pop. mean \$	Aver. Resp. gen. int. (ARGI) \$	95 % confid. Interval \$	95 % cred. interval \$	Length of ARGI \$	Length of 95 % confid. interval	Length of 95 % cred. interval
1	1997—salary and wages/ low-high usage	300 (346)	51,586	(48,595, 68,856) (**)	(52,939, 63,815)	(57,200, 60,860)	20,261	10,876	3,660
2	1997—salary and wages/ low-high usage	99 (114)	47,766	(45,787, 55,373) (**)	(45,304, 58,249) (**)	(49,690, 51,650)	9,586	12,945	1,960
3	1996—salary and wages/ low-high usage	318 (343)	48,927	(45,904, 61,613) (**)	(50,816, 59,773)	(51,990, 60,060)	15,709	8,957	8,070
4	1996—salary and wages/ low-high usage	104 (114)	44,872	(44,149, 54,092) (**)	(43,891, 56,536) (**)	(48,200, 50,100)	9,943	12,645	1,900
5	Five-year change in salary and wages/ low-high usage	275 (301)	7,339	(12,496, 24,039)	(14,947, 19,930)	(14,960, 19,650)	11,543	4,983	4,690
6	Five-year change in salary and wages/ low-high usage	93 (98)	3,088	(11,383, 21,589)	(12,205, 22,284)	(11,160, 17,320)	10,206	10,079	6,160
7	1997—interest and dividends/ low-high usage	231 (292)	2,589	(2153, 3554) (**)	(1740, 3745) (**)	(2,797, 2909)	1,401	2,005	112
8	1997—interest and dividends/ low-high usage	84 (101)	3,881	(2807, 5164) (**)	(2058, 5549) (**)	(3,906, 4,063)	2,357	3,491	157

9	1996— interest and dividends/ low–high usage	234 (289)	2,346	(1514, 2767) (**)	(1186, 2977) (**)	(2,097, 2,181)	1,253	1,791	84
10	1996— interest and dividends/ low–high usage	88 (99)	3,901	(2634, 4867) (**)	(2059, 5185) (**)	(3,677, 3,823)	2,233	3,126	146
11	Five-year change in interest and dividends/ low–high usage	219 (274)	912	(1032, 2576)	(966, 2184)	(1,769, 1,839)	1,544	1,218	70
12	Five-year change in interest and dividends/ low–high usage	76 (91)	1,124	(2099, 6250)	(156, 6930) (**)	(4,079, 4,242)	4,151	6,774	163

(**) Covers true value.

TABLE 3**ROOT MEAN SQUARED % ERROR (AVERAGED OVER 12 ITEMS)**

Median usage estimate	Bayes' estimates	Sample mean estimates	Mid-point estimate
89.7 %	127.06 % / 139.06 %	153.16 %	156.6 %

8. References

- [1] Ericson, W. A. (1969), 'A note on the posterior mean of a population mean', *Journal of the Royal Statistical Society (B)*, Vol. 31, No 2, pp. 332–334.
- [2] Heeringa, S. G., Hill, D. H. and Howell, D. A. (1995), 'Unfolding brackets for reducing item non-response in economic surveys', *Health and Retirement Study Working Paper Series*, Paper No 94-029.
- [3] Hogarth, R. (1980), *Judgment and choice*, John Wiley and Sons, Inc., New York.
- [4] Juster, F. T. and Smith, J. P. (1996), 'Improving the quality of economic data: Lessons from the HRS and AHEAD', mimeo, Survey Research Centre, University of Michigan.
- [5] Kennickell, A. B. (1997), 'Using range techniques with CAPI in the 1995 survey of consumer finances', *Survey of Consumer Finances Working Paper*, January 1997, Board of Governors of the Federal Reserve System, Washington, DC, 20551.
- [6] Lindley, D. V. (1965), *Introduction to probability and statistics — Part 2: Inference*, Cambridge University Press, Cambridge.
- [7] Marquis, K. H., and Press, S. J. (1999), 'Cognitive design and Bayesian modelling of a census survey of income recall', *Proceedings of the Federal Committee on Statistical Methodology Conference*, Washington, DC, 16 November 1999, pp.51–64 (see <http://bts.gov/fcsm>).
- [8] Metcalf, J. and Shimamura, A. P. (eds) (1994), *Metacognition: Knowing about knowing*, The NUT Press, Cambridge.
- [9] Press, S. J. (1989), *Bayesian statistics: Principles, models and applications*, John Wiley and Sons, Inc., New York.
- [10] Press, S. J. (1999), 'Respondent-generated intervals for recall in sample surveys', manuscript, Department of Statistics, University of California, Riverside, CA 92521-0138, January 1999 (<http://cnas.ucr.edu/~stat/press.htm>).
- [11] Press, S. J. and Tanur, J. M. (2000a), 'Experimenting with respondent-generated intervals in sample surveys', with discussions, in 'Survey Research at the Intersection of Statistics and Cognitive Psychology', Working Paper Series No 28, Monroe G. Sirken, Editor, National Centre for Health Statistics, January 2000, US Department of Health and Human Services, Centre for Disease Control and Prevention, pp.1–18.

- [12] Press, S. J. and Tanur, J. M. (2000b), ‘Respondent-generated interval estimation to reduce item non-response’, *Applied Statistical Science V*, Nova Science Publishers, in press (<http://cnas.ucr.edu/~stat/press.htm>).
- [13] Raiffa, H. and Schlaifer, R. O. (1961), *Applied statistical decision theory*, Graduate School of Business Administration, Harvard University.
- [14] Schlaifer, R. (1959), *Probability and statistics for business decisions*, McGraw Hill Book Co, Inc., New York.
- [15] Schwartz, L. and Paulin, G. (2000), ‘Improving response rates to income questions’, *American Statistical Association: Proceedings of the Section on Survey Research Methods*, pp. 965–969. Also, a more complete manuscript was submitted to the *Journal of Official Statistics*.
- [16] Spiegelhalter, D., Thomas, Andrew; and Nicky Best (May 1999), ‘WinBUGS Version 1.2 user manual’, MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK (<http://www.mrc-bsu.cam.ac.uk/bugs>).

Common trends in European school populations

Paola Sebastiani (*) and Marco Ramoni (**)

(*) *Department of Mathematics and Statistics, University of Massachusetts*

(**) *Children's Hospital Informatics Program, Harvard University Medical School*

Keywords: Autoregressive models, model-based clustering, Bayesian model selection, temporal data.

Abstract

This paper uses a novel Bayesian clustering method to categorise the temporal evolution of the share of population participating in tertiary/higher education in 14 European nations. The method represents time series as autoregressive models and applies an agglomerative clustering procedure to discover the most probable set of clusters describing the essential dynamics of these time series. To increase efficiency, the method uses a distance-based heuristic search strategy. This clustering method partitions the evolution of school population into three groups, thus revealing significant differences among tertiary/higher education in the 14 European nations.

1. Introduction

The time series in Figure 1 describe the evolution of the share of population enrolled in higher education in 14 nations of the European Community between 1970 and 1995. Our task is to group the 14 time series on the basis of their similarity in order to detect significant differences among European higher education trends. Data were provided by Unesco and Eurostat, via the r-cade databank, available at the URL <http://www-rcade.dur.ac.uk> (Unesco, 1997).

The method to solve this problem depends on the meaning we attach to similar time series. Throughout this paper, we will assume that time series are the realisation of stochastic processes and two or more time series are similar when the same process generates them. Thus, deciding whether two time series are similar is equivalent to deciding whether they are observations of the same process. Put in this way, the task of grouping of the time series can be solved as a clustering problem: given a batch of time series, we wish to cluster them so that each cluster contains time series generated by the same process. Particularly, we wish to solve this problem without specifying, a priori, the number of clusters. We solve this problem by using a novel Bayesian for method clustering of contributions.

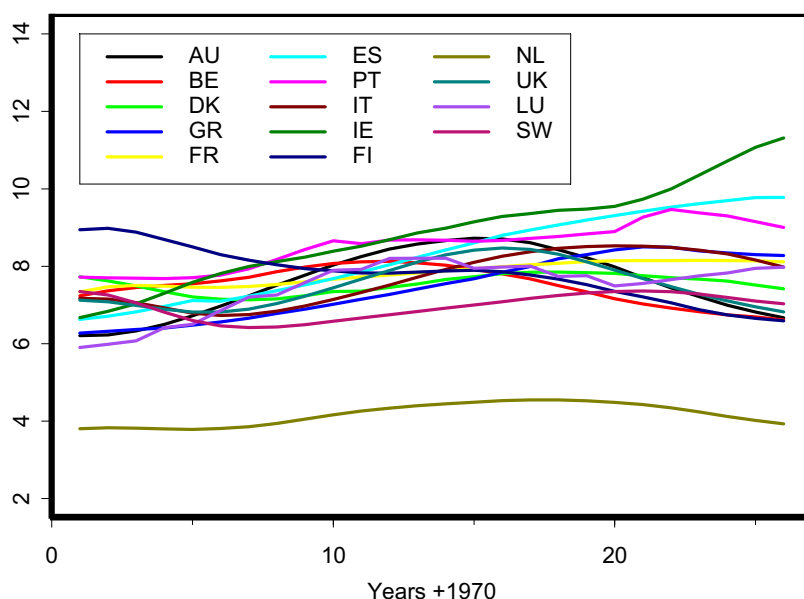
We model the stochastic process generating each time series as an autoregressive model of order p , say $AR(p)$, and then we cluster those time series that have a high posterior probability of being generated by the same $AR(p)$ model. The distinguished feature of this method is to describe a clustering of time series as a statistical model so that the clustering task can be solved as a Bayesian model selection problem. Thus, the clustering model we

look for is the most likely partition of the time series, given the data at hand and prior information about the problem.

In principle, we just need to evaluate the posterior probability of all possible clustering models of time series and select the one with maximum posterior probability. However, the number of clustering models grows exponentially with the number of time series and a heuristic search is needed to make the method feasible. The method we adopt uses a measure of similarity between $AR(p)$ models to drive the search process in a subspace of all possible clustering models. An important feature of this heuristic search is to provide a stopping rule, so that clustering can be done without assuming a given number of clusters as traditional clustering methods do.

The clustering method we use is fully described and evaluated in Sebastiani and Ramoni (2001). In the next section we briefly describe the method and the search algorithm. The analysis of the higher education data set is described in Section 3 and a discussion is in Section 4.

Figure 1: Share of population enrolled in higher education, between 1970 and 1996, in the 14 European countries: AU: Austria; BE: Belgium; DK: Denmark; GR: Greece; FR: France; ES: Spain; PT: Portugal; IT: Italy; IE: Ireland; FI: Finland; NL: The Netherland; UK: United Kingdom; Lu: Luxemburg; SW: Sweden.



2. Bayesian clustering by dynamics

The clustering method we describe here has three components: a model for the time series, the posterior probability of a clustering model and a heuristic search strategy. These three elements are described very briefly. More details are provided in Sebastiani and Ramoni (2001).

2.1. Autoregressive models

Let $S = \{y_{-p}, \dots, y_{-1}, y_1, \dots, y_t, \dots, y_n\}$ be a time series of values observed for a continuous variable Y . The series follows an $AR(p)$ model if

$$y | \beta = X\beta + \varepsilon$$

where y is the n -dimension vector $y = (y_1, \dots, y_n)$, X is the $n \times n$ matrix with t th row given by the vector of p observations y_{t-1}, \dots, y_{t-p} , $\beta = (\beta_1, \dots, \beta_p)$ is a vector of autoregressive coefficients, and ε is a vector of uncorrelated errors. We assume that the errors are normally distributed, with expectation $E(\varepsilon_t) = 0$, and variance $V(\varepsilon_t) = \sigma^2$ for any t . We shall denote by τ the precision, so that $\sigma^2 = 1/\tau$.

The value p is called the order of the autoregression, and specifies the Markov order of the series: namely that $y_t \perp (y_{-p}, \dots, y_{t-p-1}) | (y_{t-1}, \dots, y_{t-p})$, where we use the symbol \perp to denote stochastic independence. The series follows a stationary process if the roots of the polynomial $f(u) = 1 - \sum_{j=1}^p \beta_j u^j$ have moduli greater than unity. When some of the roots have moduli smaller than unity, the process is non-stationary, but typically some transformations of the data are sufficient to achieve stationarity.

The model above describes the evolution of the process around a zero mean. By adding an intercept term β_0 , the model can be extended to include a non-zero mean μ , for each y_t , so that $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ and the matrix X is augmented by a column of ones. The process mean and the autoregressive coefficients are related by: $\mu = \beta_0 / (1 - \sum_{j=1}^p \beta_j)$.

We wish to compute Bayesian estimates of the parameters β and τ . To compute the Bayesian estimates of β and τ , we need to update their joint prior density $f(\beta, \tau)$ into the posterior density $f(\beta, \tau | y)$, by using Bayes' theorem:

$$f(\beta, \tau | y) = \frac{f(\beta, \tau) f(y | \beta, \tau)}{f(y)},$$

Where $f(y | \beta, \tau)$ is the likelihood function and $f(y)$ is the marginal likelihood, which is computed as

$$f(y) = \int f(\beta, \tau) f(y | \beta, \tau) d\beta d\tau.$$

For a given autoregressive order p , we compute the likelihood function, conditional on the first p values of the time series, as

$$f(y | \beta, \tau) = \sqrt{\frac{\tau^n}{(2\pi)^n}} \exp\left(-\frac{\tau(y - X\beta)^T (y - X\beta)}{2}\right).$$

We assume as prior density for β and τ the improper prior $f(\beta, \tau) \propto X\tau^{-2}$, with $\tau > 0$ (see Jeffreys, 1946). Suppose the matrix X is of full rank, and let $\hat{\beta}$ and RSS denote respectively the ordinary least squares estimate of β and the residual sum of squares respectively:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$RSS = y^T (I_n - X(X^T X)^{-1} X^T) y$$

where I_n is the identity matrix. Then, one can show that the marginal likelihood is

$$f(y) = \frac{\left(\frac{RSS}{2}\right)^{q+2-n} \Gamma\left(\frac{n-q-2}{2}\right)}{(2\pi)^{n-q} \det(X^T X)^{1/2}}$$

where q is the dimension of the vector β . Furthermore, the posterior distribution of β and τ , is normal-gamma, with

$$\beta | y, \tau \sim N(\hat{\beta}, (\tau(X^T X))^{-1})$$

$$\tau | y \sim \text{Gamma}\left(\frac{RSS}{2}, \frac{n-q-2}{2}\right),$$

where $\text{Gamma}(a, b)$ is a gamma distribution with expected value a/b and variance a/b^2 .

Both distributions are proper whenever as the matrix X is of full rank, and $n > q + 2$. The Bayesian posterior point estimates of β and τ are \propto and $(n - q - 2) / \text{RSS}$.

2.2. Clustering

Suppose we have a batch of time series $S = \{S_1, \dots, S_m\}$, which are generated by an unknown number of stationary AR(p) models with a common autoregressive order p , and different autoregressive coefficients. We wish to cluster the time series in S according to their dynamics. Our goal is twofold:

- to find the set of clusters that gives the best partition of the data;
- to assign each time series to one and only one cluster.

Contrary to common practice, we do not want to specify, a priori, a preset number of clusters.

Formally, the clustering method regards a partition as an unobserved discrete variable C with states C_1, \dots, C_c . Each state C_k of the variable C labels, in the same way, the time series generated by the same AR(p) model and, hence, it represents a cluster of time series. The number c of states of the variable C is unknown, but it is bounded above by the total number of time series in the data set S . The clustering algorithm tries to re-label those time series that are likely to have been generated by the same AR(p) model and thus merges the initial states C_1, \dots, C_m of the variable C into a subset C_1, \dots, C_c , with $c < m$.

The specification of the number c of states of the variable C and the assignment of one of its states to each time series S_i define a statistical model M_c . This allows us to regard the clustering task as a Bayesian model selection problem, in which the model we seek is the most probable way of re-labelling time series, given the data. If $P(M_c)$ is the prior probability of each model M_c , by Bayes' theorem its posterior probability is $P(M_c | S) \propto P(M_c) f(S | M_c)$, where $f(S | M_c)$ is the marginal likelihood, now written as an explicit function of the clustering model. A model-based Bayesian solution to the clustering problem consists of selecting the clustering model with maximum posterior probability. It is shown in Sebastiani and Ramoni (2001) that, under some assumptions on the sample space, the adoption of a particular parameterisation for the model M_c and the specification of an improper-uniform prior lead to a simple, closed-form expression for the marginal likelihood $f(S | M_c)$.

Conditional on the model M_c and, hence, on a specification of the number of states of the variable C and of the labelling of the original time series, we suppose that the marginal distribution of the variable C is multinomial, with cell probabilities $\theta_k = P(C = C_k | \theta)$. Furthermore, we suppose that, conditional on $C = C_k$, the batch of m_k time series $\{S_{kj}\}$ assigned to cluster C_k are independent of the batch of time series $\{S_{lj}\}$ assigned to any other cluster C_l , and that the time series generated by the same AR(p) model in cluster C_k are mutually independent. We denote by β_k the vector of autoregression coefficients and by τ_k the precision of the AR(p) model generating the time series in cluster C_k . We suppose that each of these series can be represented as

$$y_{kj} | \beta_k, \tau_k = X_{kj} \beta_k + \varepsilon_{kj}.$$

The index k indicates cluster membership, and ε_{kj} is a vector of uncorrelated errors, which we assume to be normally distributed, with $E(\varepsilon_{kjt}) = 0$ and $V(\varepsilon_{kjt}) = \tau_k^{-1}$, for any t . The fact that series assigned to the same cluster C_k are characterised by the same vector of autoregression coefficients β_k , and by the same variance $\sigma^2_k = \tau_k^{-1}$, allows us to represent the whole batch of series $\{S_{kj}\}$ in cluster C_k as

$$y_k | \beta_k, \tau_k = X_k \beta_k + \varepsilon_k$$

where the vector y_k and the matrix X_k are defined as

$$y_k = \begin{pmatrix} y_{k1} \\ \vdots \\ y_{km_k} \end{pmatrix}$$

$$X_k = \begin{pmatrix} X_{k1} \\ \vdots \\ X_{km_k} \end{pmatrix}$$

Let β denote the set of parameter vectors $\beta = (\beta_k)$, where each β_k is a random vector, and let τ denote the set of parameters $\tau = (\tau_k)$, for $k = 1, \dots, c$. Then, by the independence of series assigned to different clusters, the overall likelihood function is

$$f(y | \theta, \beta, \tau) = \prod_{k=1}^c \theta_k^{m_k} f(y_k | X_k, \beta_k, \tau_k)$$

where m_k is the number of time series that are assigned to cluster C_k . Here, the overall likelihood is conditional on the set of $c(p+2)$ values upon which the likelihood function of each series is conditioned.

We take as our prior distribution for θ a Dirichlet $D(\alpha_1, \dots, \alpha_c)$, and assign the improper prior with density $f(\beta, \tau) \propto \prod_k \tau_k^{-2}$ to β and τ . Then, using standard results on Dirichlet integration, it is easy to show that the marginal likelihood is

$$\begin{aligned} f(y | M_c) &= \int f(y | \theta, \beta, \tau) f(\theta, \beta, \tau) d\theta d\beta d\tau \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha + m)} \prod_{k=1}^c \frac{\Gamma(\alpha_k + m_k)}{\Gamma(\alpha_k)} \frac{(RSS_k / 2)^{q+2-n_k} \Gamma((n_k - q - 2)/2)}{(2\pi)^{(n_k - q)/2} \det(X_k^T X_k)^{1/2}} \end{aligned}$$

where $\alpha = \sum_k \alpha_k$ is the overall cluster prior precision, n_k is the dimension of the vector y_k , and $RSS_k = y_k^T (I_n - X_k (X_k^T X_k)^{-1} X_k^T) y_k$ is the residual sum of squares in cluster C_k . The marginal likelihood is well defined as long as each matrix X_k is of full rank.

Once the most likely partition has been selected a posteriori, each cluster C_k is associated with the parameters β_k , which model the autoregression equation, and the precision τ_k . The posterior distribution of $\beta_k | \tau_k, y_k$ is $N(\hat{\beta}_k, (\tau_k (X_k^T X_k))^{-1})$ with $\hat{\beta}_k = (X_k^T X_k)^{-1} X_k^T y_k$, while the posterior distribution of $\tau_k | y_k$ is $\text{Gamma}((RSS_k / 2), (n_k - q - 2)/2)$. The marginal posterior distribution of the autoregression coefficients $\beta_k | y_k$ is a non-central Student's t , with expectation $\hat{\beta}_k$, which provides a point estimate of β_k . The estimate of the within cluster precision τ_k is $((n_k - q - 2) / RSS_k)$. The probability that $C = C_k$ is estimated by $\hat{\theta}_k = (\alpha_k + m_k) / (\alpha + m)$.

In our application, we use a symmetric prior distribution for the parameter vector θ , with a common prior precision α . The initial m hyper-parameters α_k are set equal to α / m and, when two time series are assigned to the same cluster C_k , their hyperparameters are summed up. Thus, the hyperparameters of a cluster merging m_k time series will be

$m_k(\alpha/m)$. In this way, the specification of the prior hyperparameters requires only the global prior precision α , which measures the confidence in the prior model. The current implementation of the algorithm assumes that the series follow stationary autoregressive models of a given order p , and then checks that the stationarity conditions are met at the end of the clustering process.

2.3. Search

In principle, the clustering method described in the previous section requires one to compute the posterior probability of each clustering model and then choose the clustering model with maximum posterior probability. Since the number of possible partitions grows exponentially with the number of series, a heuristic method is required to make the search feasible.

Our method uses a measure of similarity between $AR(p)$ models to efficiently guide the search process in a subset of all possible clustering models. Since all $AR(p)$ models have the same order, this similarity measure is an estimate of the symmetric Kullback-Liebler divergence (Jeffreys, 1946) between marginal posterior distributions of the autoregressive coefficients $\beta_k | y_k$ associated with the clusters. The estimate is given by computing the symmetric Kullback-Liebler divergence for every pair of parameters β_k, β_j , assuming a normal distribution conditional on the within-cluster precisions τ_k, τ_j . The precisions are then replaced by their posterior estimates.

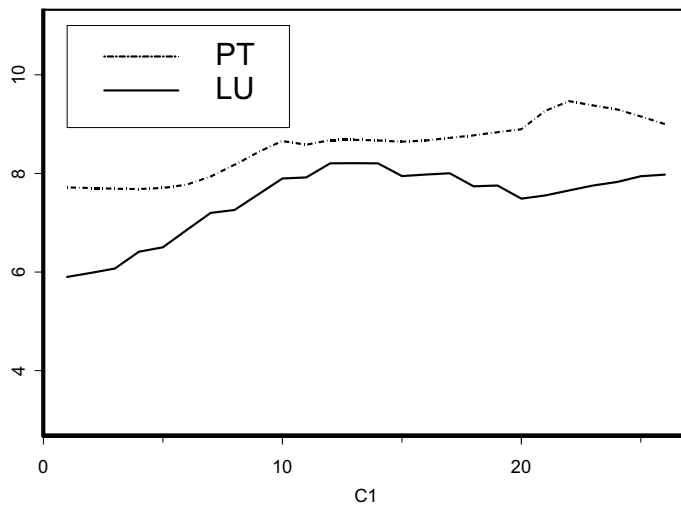
Initially, the algorithm transforms the time series in S into a set of m $AR(p)$ models, using the procedure described in the previous section, and computes the set of $m(m-1)/2$ pair-wise distances between posterior distributions of the parameters. Then, the algorithm sorts the generated distances, labels in the same way the two closest $AR(p)$ models and evaluates whether the resulting clustering model M_c , in which the two closest $AR(p)$ models are assigned to the same cluster, is more probable than the model M_s in which they are distinct. If the probability $P(M_c | y)$ is larger than $P(M_s | y)$, the algorithm updates the set of series by replacing the two series with the cluster resulting from their merging. Consequently, the algorithm updates the set of ordered distances by removing all the ordered pairs involving the merged time series, and by adding the distances between the parameters of the new $AR(p)$ model and the remaining models in the set. The procedure is then iterated on the new set. If the probability $P(M_c | y)$ is not larger than $P(M_s | y)$, the algorithm tries to merge the second best, the third best, and so on, until the set of pairs is empty and, in this case, returns the most probable partition found thus far. The rationale behind this heuristic method is that merging closest $AR(p)$ models first should speed up the search for clustering models with large posterior probability. Empirical evaluations of the methods on simulated data appear to support this intuition (see Sebastiani and Ramoni, 2001).

3. Analysis

We apply the clustering algorithm described in Section 2 to the analysis of the 14 time series reporting the temporal evolution of the share of the population engaged in tertiary/higher education in 14 European countries depicted in Figure 1. Since the average length of a university degree across European nations is three or four years, we applied the clustering algorithm under the assumption that all time series were generated by stationary AR(3) models with a non-zero mean. We assumed $\alpha = 1$, the improper prior with density $f(\beta, \tau) \propto \prod_k \tau_k^{-2}$, and uniform prior on all clustering models. Stationarity of the autoregressive models was checked at the end of the clustering process. Figure 2, 3 and 4 show the three clusters of time series found by the algorithm.

Cluster C_1 groups the evolutions of the proportion of the population enrolled in higher education institutes in Portugal and Luxembourg, see Figure 2. The estimates of the autoregression coefficients are $\hat{\beta}_0 \cong 0.657$, $\hat{\beta}_1 \cong 1.133$, $\hat{\beta}_2 \cong 0.044$ and $\hat{\beta}_3 \cong -0.254$. Thus, the model is stationary — the roots of the polynomial $f(u)$ are $-2.38, 1.28 \pm 0.11i$ — with a mean $\hat{\mu} \cong 8.532\%$. Note that the time series describing the evolution of school population in Luxembourg has a slight increasing trend during the 1970s, and then becomes stationary, with a mean slightly above 8 %.

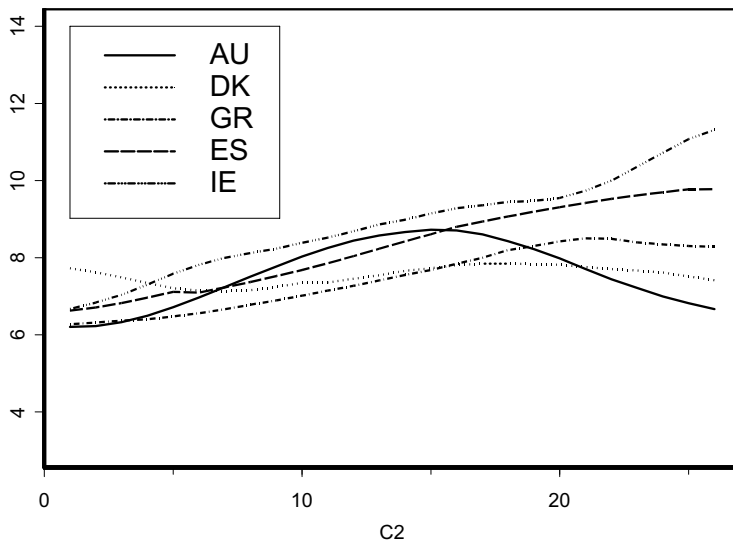
Figure 2: Cluster C1 groups the evolution of school population in Portugal and Luxembourg



The evolutions of the proportion of the population enrolled in higher education institutes in Austria, Denmark, Greece, Spain and Ireland are merged into cluster C_2 in Figure 3. The

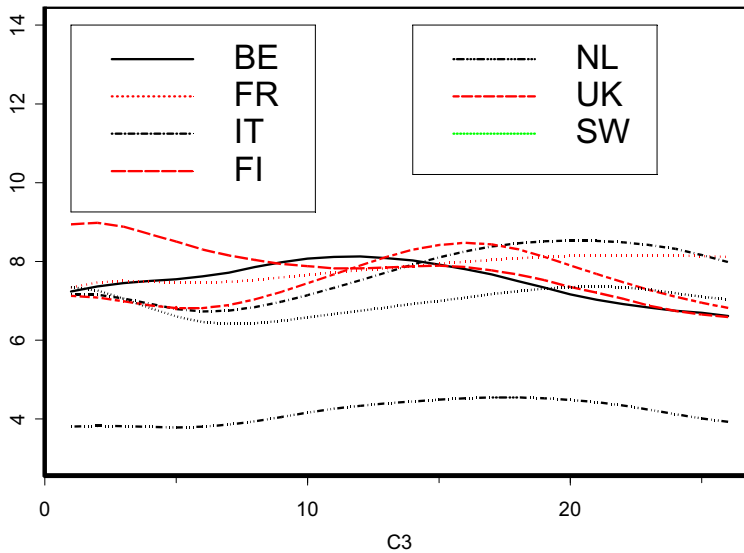
estimates of the autoregression coefficients are $\hat{\beta}_0 \cong 0.074$, $\hat{\beta}_1 \cong 2.085$, $\hat{\beta}_2 \cong -1.233$ and $\hat{\beta}_3 \cong 0.138$, with a mean $\hat{\mu} \cong 7.4$. The AR(3) model is stationary, with roots of the polynomial $f(u)$ equal to 6.09 and $1.02 \pm 0.1i$.

Figure 3: Cluster of time series describing the evolution of school population in Austria, Denmark, Greece, Spain and Ireland



Of the series assigned to this cluster, those describing the evolution of school population in Austria, Denmark and Greece are evidently stationary, while the time series describing the evolution of school population in Spain and, particularly, Ireland exhibit some trend. The assignment of the two series to this cluster could indicate that the increasing trend is only temporary, and that the proportion of the population enrolled in higher education institutes becomes stable during the 1990s.

Figure 4: Cluster merging the evolution of school population in Belgium, France, Italy, Finland, the Netherlands, the United Kingdom and Sweden



Cluster C_3 in Figure 4 groups the evolutions of the proportion of the population enrolled in higher education institutes in Belgium, France, Italy, the Netherlands, Finland, United Kingdom and Sweden. The estimates of the autoregression coefficients are $\hat{\beta}_0 \cong 0.015$, $\hat{\beta}_1 \cong 2.593$, $\hat{\beta}_2 \cong -2.283$, and $\hat{\beta}_3 \cong 0.688$, thus defining a stationary autoregression equation, with roots of the polynomial $f(u)$ equal to 1.023 and $1. \pm 0.32i$. The mean of the process is $\hat{\mu} \cong 7.5$.

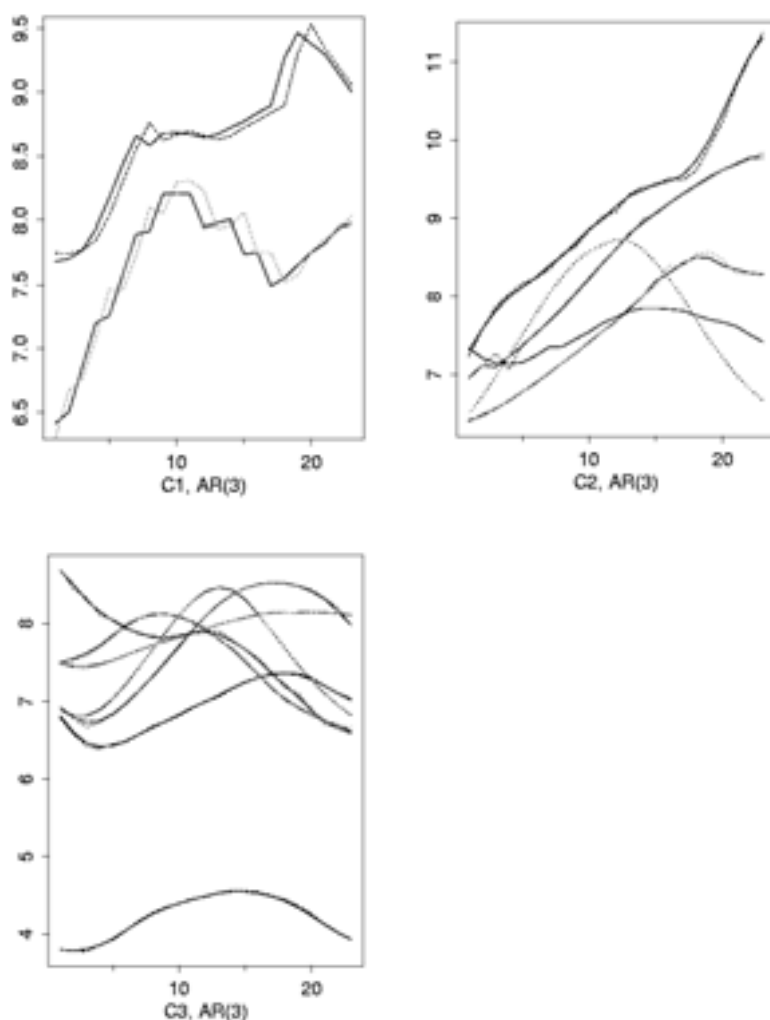
This cluster groups the European nations that have been consistently stronger from an economic point of view in the past 30 years. All these nations have a solid higher education tradition, and university curricula lasting, on the average, four years. All series assigned to this cluster are increasing up to the 1980s, and then decrease. This fact would be consistent with the large demand for highly-skilled labour and for higher education created by the pace of economic development in Europe in the 1960s. The contraction of the population together with the economic recession in the 1980s, could be responsible for the decrease of the proportion of population enrolled in higher education in the late 1980s and the 1990s.

The means of the processes generating the time series assigned to clusters C_2 and C_3 are essentially the same. However, the autoregressive equation for cluster C_3 describes a more stable process around the mean, with smaller fluctuations. Thus, the results would suggest a more stable higher education enrolment in Italy, France, the Netherlands, United

Kingdom, Belgium, Finland and Sweden, compared to Austria, Denmark, Greece, Spain and Ireland.

The fact that the time series describing the evolution of the population in higher education of the Netherlands is assigned to the third cluster is slightly disappointing: the dynamic of this series is similar to that of the other series in the cluster, but this series has a different mean. To evaluate the influence of this time series on the results, we run the clustering algorithm excluding the time series of the Netherlands. The algorithm found the same three clusters, thus showing that this series is not ‘influential’.

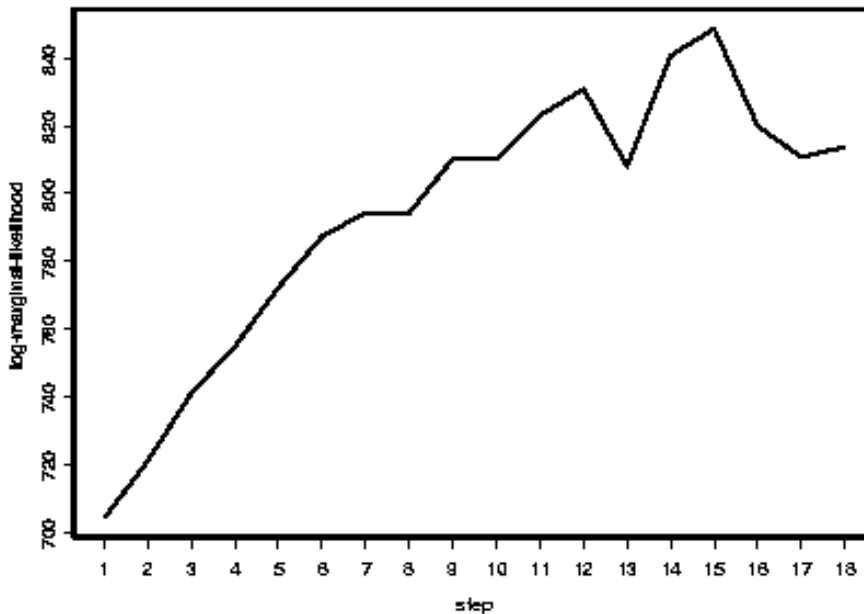
Figure 5: Observed (continuous line) and fitted (dash line) time series in the clusters in Figure 2



During the analysis we assumed the time series were generated by AR(3) models. Plots of the observed and fitted values within clusters provide an overall assessment of the robustness of the result with respect to this assumption. Figure 5 plots the time series of observed values in the three clusters and values fitted using the AR(3) models associated with each cluster. The close match supports the assumption that AR(3) models are a good approximation of the processes generating the original 14 series.

Finally, we note that the search algorithm found the three clusters of time series in just 18 steps. This number is much smaller than the total number of clusters to be considered without the heuristic search. Figure 6 shows the increase of the log-marginal likelihood — up to a constant — at each step of the agglomerative search procedure. In the first seven steps, there is a linear increase of the marginal likelihood. Thus, merging the time series that belong to the clusters with nearest autoregressive coefficients increases the marginal likelihood. In the next eight steps, merging the closest clusters does not always increase the marginal likelihood, so that the merging of the ‘second best’ is evaluated and accepted. This is so until step 15, when the algorithm has merged the 14 time series into three clusters. At this point, the three possible merging of two clusters at a time are evaluated and, since they all result in a decrease of the marginal likelihood, the algorithm stops and returns the three clusters so found.

Figure 6: Change of the marginal likelihood, in logarithmic scale, at each step of the agglomerative search procedure



4. Discussion and related work

Autoregressive models have received great attention, (see Box and Jenkins, 1976, for a systematic exposition and West and Harrison, 1997, for a Bayesian analysis). Bayesian model-based clustering was originally proposed by Banfield and Raftery (1993), to cluster static data. Ramoni et al. (2000, 2001) proposed a Bayesian clustering by dynamics algorithm, called BCD, to cluster discrete time series. BCD clusters time series modelled as Markov chains and, contrary to popular methods, finds also the number of clusters. Notwithstanding the somewhat restrictive Markov chain assumption, BCD has been applied successfully to cluster robot experiences based on sensory inputs (see Sebastiani et al., 2001), simulated war games (Sebastiani et al., 1999), as well as the behaviour of stocks in the financial market and automated learning and generation of Bach's counterpoint.

Unlike BCD, the algorithm used in this paper clusters time series of continuous variables. The different type of data requires different modelling assumptions thus producing an algorithm which is similar to BCD, in being Bayesian and model-based, but its methodology is novel. The heuristic search used by the clustering method is similar to that implemented in BCD although, here, the search is driven by a distance between posterior distributions of parameters characterising the AR(p) models of different clusters, while in BCD the search uses the distance between predictive distributions of estimated Markov chains.

The model selection strategy of our algorithm seeks the clustering model with maximum posterior probability. Other choices here would be possible such as selecting the median posterior probability model (Barbieri and Berger, 2000). One would need to compare these different model choices and see whether a similar heuristic search can be developed when the algorithm seeks for the median posterior probability model.

At first glance, modelling time series with autoregression equations of the same order may appear to be a severe restriction. We have investigated the limitation of this assumption in simulated data (see Sebastiani and Ramoni, 2001) and the emerging result is that the results of our clustering method are robust to misspecification of the autoregressive order.

5. Acknowledgements

This research was supported by Eurostat, under contract EP29105. The authors thank Ed George and an anonymous referee for their invaluable help to improve the paper.

6. References

- [1] Banfield, J. D. and Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics*, Vol. 49, pp. 803–821.
- [2] Barbieri, M. and Berger, J. O. (2000), 'Optimal predictive variable selection', ISDS Discussion Paper, Duke University.
- [3] Box, G. E. P. and Jenkins, G. M. (1976), *Time series analysis: Forecasting and control*, Holden-Day, San Francisco, CA.
- [4] Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation procedures', *Proceedings of the Royal Society, London, A*, Vol. 186, pp. 453–461.
- [5] Ramoni, M., Sebastiani, P. and Cohen, P. (2000), 'Multivariate clustering by dynamics', *Proceedings of the 17th National Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 633–638.
- [6] Ramoni, M., Sebastiani, P. and Cohen, P. (2001), 'Bayesian clustering by dynamics', *Machine Learning*, to appear.
- [7] Sebastiani, P. and Ramoni, M. (2000), 'Bayesian model-based clustering of time series', submitted.
- [8] Sebastiani, P., Ramoni, M. and Cohen, P. (2001), 'Bayesian analysis of sensory inputs of a mobile robot', *Proceedings of the Fifth Workshop on Case Studies in Bayesian Statistics*, pp. 379–395.
- [9] Sebastiani, P., Ramoni, M., Cohen, P., Warwick, J. and Davis, J. (1999), 'Discovering dynamics using bayesian clustering', *Proceedings of the Third International Symposium on Intelligent Data Analysis*, lecture notes in computer science, Springer, New York, pp. 199–210.
- [10] Unesco (1997), 'Schooling population (computer file), Paris:Unesco (producer), r-cade online service (distributor)', Universities of Durham and Essex.
- [11] West, M. and Harrison, J. (1997), *Bayesian forecasting and dynamic models*, Second edition, Springer, New York, NY.

On Bayesian record linkage

Marco Fortini (*), Brunero Liseo (**), Alessandra Nuccitelli (*)
and Mauro Scanu (*)

(*) *Italian National Statistical Institute (ISTAT), Italy*

(**) *University of Rome 'La Sapienza', Italy*

Keywords: false match rate, integration of data sources, Bayesian decision rules, MCMC

Abstract

Record linkage refers to the use of an algorithmic technique to match records from different data sets that correspond to the same statistical unit (Belin and Rubin, 1995). In this paper, we propose a fully Bayesian approach to record linkage. We use standard Metropolis-Hastings and simulated annealing algorithms to derive the marginal posterior distribution of a matrix-valued parameter which indicates the 'configuration' of matches between the two lists. We suggest using different inferential summaries of the posterior: in particular, we discuss the use of the posterior mode. Alternatively, we sketch the possibility of using a formal Bayesian decision theory approach.

1. Introduction

Record linkage refers to the use of an algorithmic technique to match records from different data sets that correspond to the same statistical unit (Belin and Rubin, 1995). The need for record linkage is ubiquitous in official statistics. For example, record linkage is a necessary preliminary step when the size of a population is estimated via capture-recapture techniques, especially when the target population is elusive (non-regular immigrants in the European Community are an example) and differences in identification variables in the two occasions are frequent. Another example, particularly important for Statistical Institutes, is given by the possibility of using administrative databases in order to complete files in a survey, relieving the response burden.

Suppose we have two computer files \mathbf{X}_A and \mathbf{X}_B whose records relate respectively to units (e.g. individuals, firms) of partially overlapping populations \mathcal{A} and \mathcal{B} and consist of several fields, or variables, either quantitative or qualitative. For example, in a file of individuals, fields can be 'surname', 'age', 'sex', etc. The objective of record linkage is to find all the pairs of units (a, b) , $a \in \mathcal{A}$ and $b \in \mathcal{B}$, such that a and b refer actually to the same unit. Hence, a record linkage procedure is a decision rule which, for each single pair of records, can take only three decisions: link, possible

link and non-link (see Fellegi and Sunter, 1969). The decision rule is based on the comparison of the common fields of the two files of records, which are denoted key (or matching) variables. We assume that the key variables jointly identify any single individual of the population. However, the key variables are likely to be observed with errors which make the linkage process not trivial.

From a classical viewpoint, the decision rule is chosen in order to minimise the expected number of no-decisions (Fellegi and Sunter, 1969). As a measure of performance of the procedure, the false match rate (FMR), defined as the number of false declared matches divided by the total number of declared matches, can be adopted. Belin and Rubin (1995) propose a method of estimation for FMR.

In this paper we propose some Bayesian strategies for record linkage. The Bayesian framework is particularly suitable for the solution of the following problems: (1) exact computation of the probability that each pair is a match, conditional on the observed data (the comparison of the key variables); (2) computation of conditional probabilities that more pairs are simultaneously matches. The first point relies on the fact that probabilities obtained by conditioning on observed events are more directly interpretable than those obtained by conditioning on unobservable hypotheses. The second represents an improvement to the classical methods, where decision rules establish separately for each pair whether they refer to the same unit or not, without considering the compatibility constraints, unless additional procedures based on operational research techniques are used (see Jaro, 1989).

In this paper we consider, as the quantity of interest, a matrix-valued parameter \mathbf{c} which represents the true pattern of matches between the two lists. We obtain an MCMC sample from the marginal posterior distribution of \mathbf{c} and we discuss several possible inferential summaries. In particular we propose the use of the posterior mode(s) as point estimate(s) of \mathbf{c} . Furthermore we briefly introduce alternative estimates; one is based on a decision-theoretic approach, using a loss function that justifies FMR as a measure of performance of record linkage techniques; the other is a more eclectic approach which slightly deviates from the Bayesian road at the reward of a more flexible analysis. There are not many papers on Bayesian analysis of record linkage and related problems. Fienberg et al. (1997) give a Bayesian model that formalises comparisons of a set of variables observed in two distinct occasions given in a disclosure problem. Larsen (1999) outlines another Bayesian approach for record linkage which uses a mixture model.

The combinatorial nature of the record linkage problems makes the analytic use of the posterior distribution of the parameters involved practically impossible; the use of a simple MCMC algorithm (basically a Metropolis-

Hastings one plus a simulated annealing optimisation step) renders the analysis computationally feasible even for blocks of units (see below) as large as 100, although we feel that dramatic improvements can be expected and we are currently working along this direction. The usual list sizes of record linkage analyses are very large. In practical applications, however, it is always the case that the entire data sets are divided into blocks, according to some reliable key variable (such as the geographic blocks described in Jaro, 1989). In fact, very small blocks will force false non-matches (missed matches), but very large blocks will increase computing and allow more false matches. The results would depend strongly on the quality of the matching information. We believe that record linkage procedures which are able to manage blocks as large as 100 units per block can be usefully applied in practice.

Throughout the paper, random variables will be denoted by capital letters and the lower case will be used for the corresponding realisations. Both matrices and vectors will be denoted by bold characters. Sets will be denoted by calligraphic characters.

2. The statistical model

Following Fellegi and Sunter (1969), let $\mathcal{A} \times \mathcal{B}$ be the set of ordered pairs of records, i.e. $\mathcal{A} \times \mathcal{B} = \{(a, b) : a \in \mathcal{A}, b \in \mathcal{B}\}$. We can split the set $\mathcal{A} \times \mathcal{B}$ into two disjoint sets: the set of matches, namely $\mathcal{M} = \{(a, b) : a = b\}$ and the set of non-matches, $\mathcal{U} = \{(a, b) : a \neq b\}$. We denote with $|S|$ the cardinality of a set S . It is worth noting that $|\mathcal{M}|$ is typically much smaller than $|\mathcal{U}|$.

Let (X_1, X_2, \dots, X_k) be the key variables observed on the two sets of units \mathcal{A} and \mathcal{B} . The corresponding design matrices are defined as $\mathbf{x}_{\mathcal{A}} = \{x_{a,j}^{\mathcal{A}}\}$ and $\mathbf{x}_{\mathcal{B}} = \{x_{b,j}^{\mathcal{B}}\}$, where $x_{a,j}^{\mathcal{A}}$ denotes the observation of the variable X_j on the a -th unit of the \mathcal{A} list, $a = 1, 2, \dots, \nu_{\mathcal{A}}$, and $x_{b,j}^{\mathcal{B}}$ corresponds to the observation of the same variable on the b -th unit of the set \mathcal{B} , $b = 1, 2, \dots, \nu_{\mathcal{B}}$, $j = 1, 2, \dots, k$.

In order to assess whether unit $a \in \mathcal{A}$ and unit $b \in \mathcal{B}$ are actually the same unit, it is crucial to give a reasonable definition of the comparison between the corresponding observed vectors, i.e. the a -th row in $\mathbf{x}_{\mathcal{A}}$ and the b -th row in $\mathbf{x}_{\mathcal{B}}$.

The comparison between two individuals (a, b) is expressed by a vector of k indicator functions $\mathbf{y}^{ab} = (y_1^{ab}, y_2^{ab}, \dots, y_k^{ab})^T$, where

$$y_j^{ab} = \begin{cases} 1 & \text{if } x_{a,j}^A = x_{b,j}^B \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

The overall comparison space will be denoted by \mathcal{D} .

Definition (1) is still applicable in the case of blank components in one or both the matrices \mathbf{x}_A and \mathbf{x}_B . One simply assumes that the indicator function y_j^{ab} is equal to 1 if and only if $x_{a,j}^A$ and $x_{b,j}^B$ are both not blank and equal.

The observed indicator functions y_j^{ab} can be considered as observations of $\nu_A \times \nu_B \times k$ random variables, each one defining the occurrence of differences between the vectors \mathbf{x}_a^A and \mathbf{x}_b^B in the two lists and their corresponding probabilities. The probability model relative to these observations can be defined as follows.

2.1. The likelihood function

Consider first the set of pairs (a, b) that correspond to true matches, i.e. the pairs in \mathcal{M} . Ideally, if the key variables were observed without error, and if individuals present in both files answered coherently in both the occasions, then the comparison vector $\mathbf{y}^{ab} = (y_1^{ab}, y_2^{ab}, \dots, y_k^{ab})^T$ should be a vector composed of 1s. However in large data sets it is very likely to have errors in the answers, typing errors and inconsistencies between answers to the same question in different times. As a consequence, one should allow the vector of comparisons between the key variables \mathbf{y}^{ab} to assume, with appropriate probabilities, all the 2^k combinations of 0's and 1's in a k -dimensional vector. In addition, we assume the random vectors \mathbf{Y}^{ab} associated to each pair $(a, b) \in \mathcal{M}$ are i.i.d. conditionally on pairs within \mathcal{M} . The most natural distribution for the \mathbf{Y}^{ab} s is the multinomial one, i.e.

$$P(\mathbf{Y}^{ab} = \mathbf{i} | c_{a,b} = 1, \mathbf{m}) = m_{i_1, i_2, \dots, i_k} = m_{\mathbf{i}} \quad \mathbf{i} \in \mathcal{D}, \quad (2)$$

where $c_{a,b}$ is the indicator of the membership of the pair (a, b) to \mathcal{M} .

In this notation, the matrix $\mathbf{c} = \{c_{a,b}\}$ becomes the actual parameter of interest in a record linkage procedure. We assume that the matrix \mathbf{c} can take values on the set \mathcal{C} of all the matrices satisfying the following obvious consistency conditions:

$$\begin{cases} c_{a,b} \in \{0, 1\} & \forall a \in \mathcal{A}, b \in \mathcal{B} \\ \sum_{b=1}^{v_B} c_{a,b} \leq 1 & \forall a \in \mathcal{A} \\ \sum_{a=1}^{v_A} c_{a,b} \leq 1 & \forall b \in \mathcal{B} \end{cases} \quad (3)$$

It is worth noting that imposing constraints on a parameter space is straightforward in a Bayesian approach. On the other hand, this is very complicated from a frequentist perspective. Consequently, currently used record linkage procedures must complete the statistical data analysis with a reallocation procedure which eliminates inconsistencies among the results of different tests (see Jaro, 1989, and the problem posed by Larsen, 1999, paragraph 3.3): this is automatic in a Bayesian framework.

The same argument used for (2) can be applied to all the pairs in \mathcal{U} . We define: $P(\mathbf{Y}^{ab} = \mathbf{i} | c_{a,b} = 0, \mathbf{u}) = u_{i_1, i_2, \dots, i_k} = u_{\mathbf{i}}, \mathbf{i} \in \mathcal{D}$, where the random comparison vectors \mathbf{Y}^{ab} are still multinomial, with a different set of parameters, \mathbf{u} .

The likelihood function for \mathbf{m} , \mathbf{u} , and \mathbf{c} associated with the observed matrix \mathbf{y} is:

$$\begin{aligned} L(\mathbf{c}, \mathbf{m}, \mathbf{u} | \mathbf{y}) &= \prod_{a=1}^{v_A} \prod_{b=1}^{v_B} \left(\prod_{\mathbf{i} \in \mathcal{D}} m_{\mathbf{i}}^{d(\mathbf{y}^{ab}, \mathbf{i})} \right)^{c_{a,b}} \left(\prod_{\mathbf{i} \in \mathcal{D}} u_{\mathbf{i}}^{d(\mathbf{y}^{ab}, \mathbf{i})} \right)^{1-c_{a,b}} = \\ &= \prod_{\mathbf{i} \in \mathcal{D}} m_{\mathbf{i}}^{\sum_{a,b} d(\mathbf{y}^{ab}, \mathbf{i}) c_{a,b}} u_{\mathbf{i}}^{\sum_{a,b} d(\mathbf{y}^{ab}, \mathbf{i}) (1-c_{a,b})} \end{aligned} \quad (4)$$

where, for every $\mathbf{i} \in \mathcal{D}$,

$$d(\mathbf{y}^{ab}, \mathbf{i}) = \begin{cases} 1 & \text{if } \mathbf{y}^{ab} = \mathbf{i} \\ 0 & \text{otherwise} \end{cases}.$$

The likelihood function (4) could be also analysed from a mixture model perspective, with a fixed number of components (matches, non-matches); see Winkler (1994, 1995), Larsen (1999), Larsen and Rubin (2000) and references therein.

The likelihood (4) is more than saturated and consequently cannot be managed in a classical framework unless severe restrictions are posed on the dependence structure among the comparison variables Y_j^{ab} , $j=1, 2, \dots, k$, (see for instance Fellegi and Sunter, 1969, Jaro, 1989, Winkler, 1993). In a Bayesian framework this is done by modelling the (conditional) prior distribution of \mathbf{m} and \mathbf{u} .

2.2. The prior distributions

The likelihood function (4) depends on two sets of parameters: the parameters of interest (the matrix \mathbf{c}) and two vectors of nuisance parameters (\mathbf{m} and \mathbf{u}). In real applications, for both sets of parameters it is reasonable and typical to have some kind of information which should enter the analysis in terms of suitable prior distributions. From now on, we will consider \mathbf{c} , \mathbf{m} and \mathbf{u} as random objects whose prior distributions are discussed in the following.

The prior distribution for the random matrix \mathbf{C} can be given in two steps. The first step consists of a prior distribution $\pi_H(h)$, $h=0, 1, \dots, \nu_A \wedge \nu_B$, on the number of matched pairs H in the two lists. This is usually the step where the researcher can collect information easily, looking at previous experiences or at the statistical characteristics of the data sets (e.g. if the two data sets refer respectively to a census and a sample, we can expect a large number of matched pairs). The second step consists of a conditional distribution of the configurations \mathbf{C} given the number of matches. The prior distribution for \mathbf{C} is defined by the relationship:

$$P(\mathbf{C} = \mathbf{c}) = \pi_H(h)P(\mathbf{C} = \mathbf{c} | H = h), \quad (5)$$

where the first equality holds because $P(\mathbf{C} = \mathbf{c}) = P(\mathbf{C} = \mathbf{c}, H = h)$, for suitable h . Here we propose a reasonable and natural prior distribution for \mathbf{C} , according to (5).

Define C^h as

$$C^h = \left\{ \mathbf{c} \in \mathbf{C} : \sum_{a,b} c_{a,b} = h \right\} \quad h = 0, 1, \dots, \nu_A \wedge \nu_B.$$

Then we adopt a uniform conditional prior distribution for $\mathbf{C} | (H = h)$ over C^h . Different priors can be chosen for H . Here, for convenience, we adopt a binomial distribution with parameters $\nu_A \wedge \nu_B$ and ξ .

The latter should reflect our beliefs on the most probable number of matches (e.g. ξ could be calibrated as an average relative frequency of

observed matches in previous and similar applications), while the former is a flat distribution since, in general, we do not possess, a priori, information able to distinguish between different pairs. However both the previous prior distributions can be reasonably modified in order to take into account additional and more specific information (Larsen and Rubin, 2001).

Although not of primary interest, the nuisance parameters \mathbf{m} and \mathbf{u} play an important role in the model. Previous approaches to record linkage developed a complex machinery (involving the EM algorithm and strict and often arbitrary hypotheses on the statistical model) in order to obtain reasonable estimates of \mathbf{m} and \mathbf{u} . In a Bayesian view, one simply integrates out the nuisance parameters from the likelihood (4), after having specified a suitable prior (conditional on \mathbf{c}).

In what follows, we will assume that the random vectors \mathbf{M} and \mathbf{U} are a priori independent of \mathbf{C} and, for computational reasons, we assume that \mathbf{M} and \mathbf{U} follow a Dirichlet distribution: $\mathbf{M} \sim \mathcal{D}_{|\mathcal{D}|-1}(\cdot; \alpha)$ and $\mathbf{U} \sim \mathcal{D}_{|\mathcal{D}|-1}(\cdot; \beta)$, where $\alpha_i > 0$ and $\beta_i > 0$, $\forall \mathbf{i} \in \mathcal{D}$. In addition, we also assume a priori independence between the r.v.'s \mathbf{M} and \mathbf{U} , i.e. independence between the available information on the two different multinomial parameters \mathbf{m} and \mathbf{u} . The calibration of the hyperparameters is crucial, and deserves some comments.

In our experiments, we choose the following parameterisation:

$$\alpha_i = \theta \sum_{j=1}^k i_j^{-\phi} \quad \mathbf{i} \in \mathcal{D}, \theta > 0 \quad \phi \in \mathbb{R}. \quad (6)$$

Similarly we set

$$\beta_i = \theta^{-\phi} \sum_{j=1}^k i_j \quad \mathbf{i} \in \mathcal{D}, \theta > 0 \quad \phi \in \mathbb{R}. \quad (7)$$

This choice models our beliefs on the informative power of the comparison variables. In fact, the hyperparameters (6) hierarchically order the possible observations $\mathbf{i} \in \mathcal{D}$ in such a way that the prior distribution for \mathbf{M} puts more mass around 'large' values of m_i for those \mathbf{i} s with a large number of 1s. The opposite argument holds for the hyperparameters (7).

In particular it can be easily shown that, by introducing the hyperparameters as in (6) and (7), the marginal prior means of the m_i s and the u_i s are simple functions of θ only, whereas their variances depend on both θ and ϕ . The hyperparameters in (6) and (7) have also direct effects

on the statistical relationship among the comparison variables. For instance standard arguments prove that the correlation between two comparison variables has a null expected value for any θ and ϕ , whereas its variance depends on both. These considerations can guide the calibration process of the hyperparameters.

We want to stress the fact that the choice of the prior distributions for \mathbf{m} and \mathbf{u} , instead of inducing one particular association among comparison variables within the match and non-match groups (as in the EM case), defines a distribution over all the possible dependence structures. We expect that this situation makes our model more flexible; however, deeper investigations are needed in order to explore the connections among prior distributions for \mathbf{m} (and \mathbf{u}) and dependence structure on the comparison variables.

3. Posterior analysis

The likelihood function for the parameter of interest \mathbf{C} can be obtained analytically by integrating out the nuisance parameters \mathbf{m} and \mathbf{u} . Standard use of Dirichlet integrals and the fact that, for all (a, b) , $\sum_{\mathbf{i}} d(\mathbf{y}^{ab}, \mathbf{i}) = 1$, provide the following expression

$$L(\mathbf{c}|\mathbf{y}) \propto \frac{\prod_{\mathbf{i} \in \mathcal{D}} \Gamma\left(\sum_{a,b} [d(\mathbf{y}^{ab}, \mathbf{i}) c_{a,b}] + \alpha_{\mathbf{i}}\right) \Gamma\left(\sum_{a,b} [d(\mathbf{y}^{ab}, \mathbf{i}) (1 - c_{a,b})] + \beta_{\mathbf{i}}\right)}{\Gamma\left(h + \sum_{\mathbf{i} \in \mathcal{D}} \alpha_{\mathbf{i}}\right) \Gamma\left(\nu_A \times \nu_B - h + \sum_{\mathbf{i} \in \mathcal{D}} \beta_{\mathbf{i}}\right)}.$$

The unnormalised posterior distribution for \mathbf{C} is then given by

$$P(\mathbf{C} = \mathbf{c}|\mathbf{y}) \propto \pi_H(h) P(\mathbf{C} = \mathbf{c}|H = h) L(\mathbf{c}|\mathbf{y}). \quad (8)$$

Note that the integrated likelihood function and, a fortiori, the posterior distribution explicitly depend on h , the true number of matches.

3.1. The posterior distribution

We now discuss how to use the posterior distribution (8) in a record linkage analysis: this problem is in fact peculiar enough to suggest alternative strategies to obtain inferential summaries of the marginal posterior distribution for the parameter of interest. In a record linkage analysis the whole matrix \mathbf{C} is the parameter of interest and the problem is rather complicated, due to the fact that there are no simple ways to synthesise information on \mathbf{C} from the posterior distribution; in particular there is not a

natural ordering among the possible values of \mathbf{C} . Once we know the posterior distribution of \mathbf{C} , it is not clear for example how to produce a point estimate (but remember that the practical goal of a record linkage problem is exactly to produce a point estimate of \mathbf{C} !). In fact, the posterior mean does not seem reasonable here while the posterior median is even hard to define. The posterior mode is certainly more appealing, although it typically suffers from a certain degree of sensitivity to prior distribution, especially when ν_A and ν_B are large. More formally one should construct a specific loss function in order to minimise the posterior expected loss. Here we mainly consider the posterior mode as our point estimate. In the next section we will suggest alternative approaches which can be easily implemented and which will be explored elsewhere.

To illustrate the implementation of the methodology just described, we conducted a simulation study via perturbation of real data. The following application is based on individual data collected in October 1998 in the context of the test for the 2001 Italian Census of Population and Housing. Key variables are sex, day of birth, month of birth, year of birth, marital status, occupational status, relationship to head of household, and highest educational qualification. We have considered two data sets with a known number of matches and then we have deliberately ‘introduced’ errors at various extents according to a ‘completely-at-random’ mechanism. Two different error rates used for generating the perturbed data were selected on the basis of the available information from previous and similar surveys; the two error rates (respectively column 1 and 2 in Table 1) represent different quality standards we can expect from the key variables.

Table 1: Total error rates introduced in each file

Key variables	Low degree of distortion	High degree of distortion
<hr/>		
Sex	0.02	0.04
Day of birth	0.04	0.08
Month of birth	0.03	0.06
Year of birth	0.03	0.06
Marital status	0.06	0.12
Occupational status	0.24	0.48
Relationship to head of household	0.04	0.08
Highest educational qualification	0.22	0.44

As ξ is felt to be the most critical prior choice, we have considered in the simulations two different values of ξ , namely $\xi_1 = 0.6$ and $\xi_2 = 0.9$.

In the spirit of the considerations on the calibration of the hyperparameters given in the previous section, we choose $\theta = 2$ and $\phi = k / 2 = 4$, where k is the number of key variables. In a real application, a hierarchical Bayesian analysis which includes genuine prior information for the hyperparameters, ξ , θ and ϕ can also be performed. We will explore these issues elsewhere.

We used a Metropolis-Hastings algorithm to generate a sample from the posterior distribution and a simulated annealing routine to find the posterior mode(s) (details and computer codes are available upon request). This procedure has been applied in four distinct cases according to a combination of the input characteristics described previously (see the first part of Table 2). Each single list has been partitioned into four blocks with dimensions ranging from 79 to 100. In Table 2 we summarise results on the goodness of the posterior mode estimate (for each pair of blocks being compared, f represents the fraction of units in the smallest block which are actually matches; CMR is the correct match rate, i.e. the ratio of the observed number of true links and the number of matches).

Table 2: Input and output characteristics of four examples

Input characteristics	Case 1	Case 2	Case 3	Case 4
Degree of distortion	Low	High	Low	High
ξ	0.60	0.60	0.90	0.60
f	0.76	0.80	0.96	0.96
Output parameters				
CMR	0.93	0.56	0.99	0.98
FMR	0.03	0.12	0.01	0.01

Results are encouraging. In almost all the cases the FMR is remarkably low; also, in all the simulations the posterior probability of the true matrix configuration \mathbf{c}^* is close to the posterior modal probability. Further analysis is however necessary to verify the sensitivity to other prior inputs (i.e. the hyperparameters α and β).

We want to stress again the fact that much work has yet to be done, both from a methodological and a computational perspective. On the former side we are currently studying a way of relaxing the multinomial assumptions over the parameters \mathbf{m} and \mathbf{u} , by using a non-parametric Polya-tree distribution. On the latter side, heavier computations, with a larger range of possible values of the hyperparameters are necessary to check the practical implementation of our approach.

3.2. Alternative approaches

In this section we describe two alternative strategies to produce a point estimate of \mathbf{C} . First, we suggest a conservative approach which aims at overcoming the fact that the posterior mode can be, in a certain sense, a too crude estimator. This is due to several reasons; for example the posterior distribution could be very sparse, with a relatively large number of modes. Also, the determination of the mode may be very sensitive to the prior inputs.

An alternative approach can be based on a re-calibration of the posterior through the following steps:

Step 1: Find the posterior mode of H , say h^* .

Step 2: Determine: $\hat{\mathbf{c}}^{h^*} = \arg \max_{\mathbf{c}^{h^*}} P(\mathbf{C} = \mathbf{c} | \mathbf{y})$, where $\mathbf{C}^{h^*} = \{\mathbf{c} \in \mathbf{C} : \sum_{a,b} c_{a,b} = h^*\}$.

Step 3: Compare $\hat{\mathbf{c}}^{h^*}$ with other values of \mathbf{c} in a ‘neighbourhood’ $I(\hat{\mathbf{c}}^{h^*})$ of $\hat{\mathbf{c}}^{h^*}$. Here a neighbourhood may consist of all the matrices \mathbf{c} which can be obtained from $\hat{\mathbf{c}}^{h^*}$ with the deletion or addition of one match. Since neighbourhood comparison involves matrices with different numbers of matches, we propose to use the Bayes factor, in order to minimise the sensitivity to the prior.

Step 4: Select, as the point estimate, the value of \mathbf{c} which maximises the integrated likelihood $L(\mathbf{c} | \mathbf{y})$ over $I(\hat{\mathbf{c}}^{h^*})$.

The last two steps of the above algorithm can be iterated until stability has been reached. Note that this calibration step has no cost from a computational viewpoint since $L(\mathbf{c} | \mathbf{y})$ is given in closed form and the Bayes factor of a given $\mathbf{c}^{(1)}$ against $\mathbf{c}^{(2)}$ is simply given by the ratio $L(\mathbf{c}^{(1)} | \mathbf{y}) / L(\mathbf{c}^{(2)} | \mathbf{y})$. In the actual implementation of the above algorithm one should use a prior distribution for H which, in some sense, overestimates the number of matches. This way we start the algorithm with a value of H which is likely to be larger than the true value. Consequently, the point estimate obtained in step 2, hopefully ‘contains’ all the true matches. Then in the last steps of the algorithm, we try to eliminate the false matches.

A second, more formal Bayesian approach is based on the use of a loss function to produce a final decision. In the record linkage context the action space for selecting a point estimate of the matrix \mathbf{c} is naturally given by the parameter space \mathcal{C} . A popular measure of performance in the record linkage literature is given by FMR. However, FMR considers only one kind of mismatches because it does not take into account the percentage of false unmatched units. Also, the FMR can be evaluated, in real applications, only after ‘the urn has been opened’; in other words heavy clerical work is necessary to compute the FMR. To overcome this problem, Larsen and Rubin (2000) propose, as a measure of performance, a sort of posterior expected value of the FMR, although their analysis cannot be considered fully Bayesian. We put forward these ideas by defining a loss function which is based on FMR and the false unmatched rate (FUR) also. More formally, let \mathbf{c}^* be the true configuration value and let $I_{\mathcal{B}}(\cdot)$ be the indicator function of the set \mathcal{B} . We define the loss function of taking the decision $\mathbf{c} \in \mathcal{C}$ as follows:

$$\mathcal{L}(\mathbf{c}, \mathbf{c}^*) = \text{FMR} + \text{FUR} = \frac{\sum_{a,b} c_{a,b} I_{\{c_{a,b}^*=0\}}(c_{a,b})}{\sum_{a,b} c_{a,b}} + \frac{\sum_{a,b} (1 - c_{a,b}) I_{\{c_{a,b}^*=1\}}(c_{a,b})}{\nu_A \times \nu_B - \sum_{a,b} c_{a,b}}.$$

It is easily seen that the posterior expected loss associated with a decision \mathbf{c} is

$$E^{P(\mathbf{c}^*|\mathbf{y})} \mathcal{L}(\mathbf{c}, \mathbf{c}^*) = \frac{\sum_{a,b} c_{a,b} P(C_{a,b} = 0|\mathbf{y})}{\sum_{a,b} c_{a,b}} + \frac{\sum_{a,b} (1 - c_{a,b}) P(C_{a,b} = 1|\mathbf{y})}{\nu_A \times \nu_B - \sum_{a,b} c_{a,b}}.$$

Denoting with $\mathcal{F}_i = \{(a,b) : c_{a,b} = i\}$, $i = 0, 1$, it can be proved that the posterior expected loss of \mathbf{c} is given by

$$W(\mathbf{c}) = \frac{\sum_{(a,b) \in \mathcal{F}_1} P(C_{a,b} = 0|\mathbf{y})}{h} + \frac{\sum_{(a,b) \in \mathcal{F}_0} P(C_{a,b} = 1|\mathbf{y})}{\nu_A \times \nu_B - h}, \quad (9)$$

where h is the number of matches in \mathbf{c} . The minimisation of (9) can be faced, for instance, with an algorithm able to discover for each $h = 0, 1, \dots, \nu_A \wedge \nu_B$ the partial optimal solution for the subset \mathcal{C}^h . The configuration with lowest posterior expected loss (9) in \mathcal{C}^h is obviously the one with lowest $\sum_{(a,b) \in \mathcal{F}_1} P(C_{a,b} = 0|\mathbf{y})$. If $h = 0$ the solution is trivial; for fixed $h \geq 1$, this problem can be formulated in terms of the following assignment problem (see Lawler, 1976, Chapter 8): find the constants $c_{a,b}$ such that

$$\sum_{a,b} c_{a,b} P(C_{a,b} = 1 | \mathbf{y}) = \sum_{(a,b) \in F_1} P(C_{a,b} = 1 | \mathbf{y})$$

is maximum, subject to the constraints

$$\begin{aligned} \sum_{b=1}^{\nu_B} c_{a,b} &\leq 1 & a = 1, 2, \dots, \nu_A \\ \sum_{a=1}^{\nu_A} c_{a,b} &\leq 1 & b = 1, 2, \dots, \nu_B \\ c_{a,b} &\geq 0 & a = 1, 2, \dots, \nu_A, \\ & & b = 1, 2, \dots, \nu_B \\ \sum_{a,b} c_{a,b} &= h. \end{aligned}$$

This assignment problem can be solved, for instance, via the Hungarian Method, that finds the optimal configuration for each h , $h = 0, 1, \dots, \nu_A \wedge \nu_B$ (Lawler, 1976, p. 204).

Denoting with \mathbf{c}^h the configuration with lowest posterior expected loss in C^h , the global optimal solution is

$$\mathbf{c}^* = \arg \min \{W(\mathbf{c}^h), h = 0, 1, \dots, \nu_A \wedge \nu_B\}.$$

4. Acknowledgements

The authors are particularly grateful to Raffaella Succi for her comments and suggestions on the use of operational research techniques in Section 3. B. Liseo's research was supported by MURST (Italy).

5. References

- [1] Belin, T. R. and Rubin, D. B. (1995), 'A method for calibrating false-match rates in record linkage', *Journal of the American Statistical Association*, Vol. 90, pp. 694–707.
- [2] Fellegi, I. P. and Sunter, A. B. (1969), 'A theory of record linkage', *Journal of the American Statistical Association*, Vol. 64, pp. 1183–1210.
- [3] Fienberg, S. E., Makov, U. E. and Sanil, A. P. (1997), 'A Bayesian approach to data disclosure: Optimal intruder behaviour for continuous data', *Journal of Official Statistics*, Vol. 13, pp. 75–89.

- [4] Jaro, M. A. (1989), ‘Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida’, *Journal of the American Statistical Association*, **Vol. 84**, pp. 414–420.
- [5] Larsen, M. D. (1999), ‘Multiple imputation analysis of records linked using mixture models’, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 65–71.
- [6] Larsen, M. D. and Rubin, D. B. (2001), ‘Iterative automated record linkage using mixture models’, *Journal of the American Statistical Association*, Vol. 96, pp. 32–41.
- [7] Lawler, E. L. (1976), *Combinatorial optimisation: Networks and matroids*, Holt, Rinehart & Winston, New York.
- [8] Winkler, W. E. (1993), ‘Improved decision rules in the Fellegi-Sunter model of record linkage’, *Proceedings of the sectional survey research methods, American Statistical Association*, pp. 274–279.
- [9] Winkler, W. E. (1994), ‘Advanced methods for record linkage’, *Proceedings of the sectional survey research methods, American Statistical Association*, pp. 467–472.
- [10] Winkler, W. E. (1995), ‘Matching and record linkage’, in Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge M. J. and Kott, P. S. (eds), *Business survey methods*, Wiley, New York, pp. 355–384.

Note to authors

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Further information:

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

Executive Editor

ROS, Eurostat, Room A2/162a

BECH Building

L-2920, LUXEMBOURG

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

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