WP. 21
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (iii) Confidentiality aspects of statistical information taking into account register-based data

STATISTICAL CONFIDENTIALITY IN LONGITUDINAL LINKED DATA: OBJECTIVES AND ATTRIBUTES

Invited Paper

Submitted by the University of Alicante, Spain[1]

[1] Prepared by Mario Trottini, mario.trottini@ua.es.

Statistical Confidentiality in Longitudinal Linked Data: Objectives and Attributes

Mario Trottini*

* Departamento de Estadística e I.O., University of Alicante, Spain. (mario.trottini@ua.es)

Abstract. Researchers and practitioners interested in applications of multiple objective decision theory generally agree that the task of structuring the objectives and identify suitable attributes is the most important step in any formal analysis of a complex decision problem. In this paper we briefly review some of the relevant literature on the topic and we argue on its relevance for the current research on Data Disclosure Limitation, with particular emphasis on the dissemination of longitudinal linked data.

1 Introduction

A number of European and U.S. official statistical agencies have undertaken initiatives to develop longitudinal linked data sets. These are defined as microdata that contain observations from two or more related sampling frame, with measuraments for multiple time periods for all units of observation (see Abowd and Woodcock 2001, 2004). They are obtained integrating (through record linkage techniques) existing (administrative, or survey) microdata possibly collected by different agencies. Longitudinal linked data are essential to a wide range of research efforts and provide exceptionally rich source of information to address complex policy issues in key areas such as health care, education, and economic, to mention just a few. From the perspective of an official statistical agency, in addition, longitudinal linked data have great potential to enhance existing official statistics, improve data accuracy and reduce data collection redundancies (see Mackie and Bradburn 2000). The information content of longitudinal linked data, however, makes them vulnerable to disclosure. Requirements designed to protect confidentiality in the native data sets, i.e. those data sets from which the links were made, can drastically reduce the potential research (and non-research) benefits of linking.

This makes dissemination of longitudinal linked data a complex decision problem. An ideal data dissemination procedure, in fact, should: (i) allow legitimate data users to perform the statistical analysis of interest *as if* they were using the data set originally collected; (ii) reduce the risk of misuses of the data by potential intruders aimed to disclose confidential information about individual respondents, harm the data providers, or embarrass the statistical agency; (iii) be operational (it should be possible for the agency to implement the data dissemination procedure given the agency's resources - budget, time, people skill, technology etc.). This identify three conflicting objectives (that we call "maximize usefulness", "maximize safety" and "minimize cost") that no data dissemination procedure can fully achieve simultaneously. Improvement in an arbitrary subset of these objectives usually requires to reduce achievement in some of the objectives in the

complementary set and there is no data dissemination procedure which is obviously the best. In addition the above objectives are too ambiguous to be of operational use and there is no obvious measure that can be used to quantify the extent to which they are achieved by different candidates for data dissemination.

The research literature and current practice in Data Disclosure Limitation, have addressed these issues only in part and to a different extent. *Decision theory*, we believe, might provide a suitable framework to think about these problems. Within this framework a sensible choice of the data dissemination procedure requires the agency/ies responsible for it to:

a) Identify a set of suitable alternatives (candidate data dissemination procedures);

b) Defining the fundamental *objectives* in more operational terms;

c) Define suitable *attributes* that can measure the extent to which objectives are achieved when an arbitrary alternative is considered;

d) Assess the trade-off between the fundamental objectives of the problem. This means that for any arbitrary subset of the objectives the agency has to make a decision about how much of those objectives is willing to sacrifice in order to improve achievement in the others.

The research literature on decision theory have proposed guidelines for the implementation of the four steps decision analysis described above. In this paper we briefly review the most relevant results for structuring objectives and defining suitable attributes (points (b)-(c) above) and we argue on their relevance for increasing the values of existing research efforts in statistical confidentiality[1].

Section 2 reviews current alternatives for data dissemination and proposes a broader definition of "alternative". Strategies for a suitable structuring of the objectives and their relevance for the dissemination of (longitudinal linked) microdata are described in section 3. Section 4 deals with the problem of attributes definition and attributes selection. Section 5 summarizes the main results of the paper.

2 Identifying the alternatives

The research literature on statistical confidentiality has identified three ways of disseminating (longitudinal linked) microdata: (1) releasing a *masked* version of the data obtained through a suitable transformation of the original data; (2) *restricting access* by reducing the set of users or the modality of access to the data; and (3) generate *synthetic data* through multiple imputation (or other) methods.

Strengths and limitations of these approaches have been extensively studied (see, for example Doyle et al. 2001 and Abowd and Woodcock 2004). The problem of selecting an alternative for data dissemination is often reduced to compare different instances of (1) to

[1]The paper complements a companion paper presented by the author at the UNECE-EUROSTAT workshop held in Luxembourg on April 2003. In that paper (see Trottini 2003) the author discussed the trade-off problem (point (d) above) assuming as given attributes and objectives.

(3) the discussion focusing on whether broader access (data masking) is more important than greater data details (restricted access) and the ability of imputation method (syntetic data) to accomplished both for targeted complex analysis.

Despite the fact that several official statistical agencies disseminate longitudinal linked data using a combination of these three basic approaches (see for example Abowd and Lane 2003a), identification of alternatives is still understood as a comparison of the three basic approaches. We argue here that the value of each method would increase if we start thinking about data dissemination procedures as combinations of them. This idea has several rationales.

First of all, it is well recognized that data users and data users needs are very diverse. Thus a data dissemination procedure that relies only on one approach is unsatisfactory since it would likely produce a data set of insufficient detail for the more sophisticated data users while perhaps unnecessarily disclosing information not needed for more basic research (Mackie and Bradburn 2000). This is specially true for longitudinal linked data that are essential both for complex modeling (often of nonlinear relationships) and for the production of official statistics targeted to a much broader and less specialized audience.

In addition the definition of data dissemination as a combination of different dissemination modes, forces the agency to think of the problem as an optimal portfolio problem. The amount of resources devoted to each dissemination mode being the "parameter" that differentiates candidate alternatives. As noted by Abowd and Lane (2003b) "Any two protection methods are correlated in their risk of disclosure of confidential information, but not perfectly. Combining the two methods can, then, produce greater data utility for any level of disclosure risk in exactly the same way that an investor can achieve greater expected return for any given level of investment risk by combining the risky assets into a portfolio".

3 Structuring the Objectives

In complex decision problems the fundamental objectives are usually too broad and ambiguous to be of operational use. A useful strategy is then to divide an objective in *lower level objectives* that clarify the interpretation of the broader objective. This "splitting" process can be repeated several times, if necessary. The stopping point depends on several considerations. First of all at each split we should make sure that the set of the lower level objectives that is produced represents all the relevant aspects of the broader objective that has been split. This constraint being satisfied, the number of lower level objectives should be kept as small as possible. Test of importance can be used to minimize the number of lower level objectives (see Keeney and Raiffa 1976, chapter 2). The splitting process generates a *hierarchy*. On the top of the hierarchy there are the fundamental objectives and at the bottom all the lowest level objectives that specify all aspects that matter to assess achievement of the fundamental objectives. The more we split an objective the easier is to define "objective" attributes for the lowest level objectives and thus an "objective" representation of the fundamental objective. However continuing splitting increases the dimension of the attribute that represents the fundamental objective and makes harder to formalize sensible trade-offs for different alternatives (since the number of elements that

are involved in the trade-off increases as well). A good compromise is to build a hierarchy as detailed as possible in order to have a representation of all the relevant aspects of the problem but use the extended hierarchy as a qualitative tool to define quantitative attributes only at higher levels. The idea of the hierarchy, in its simplicity, provides a great tool to clarify what really matter in the decision analysis. Building a hierarchy can be very helpful in decision problems where several decision makers have to reach a joint decision (this is often the case in the dissemination of longitudinal linked data). Different decision makers can merge their hierarchies and the merged hierarchy will provide a suitable framework for comparison and constructive criticisms.

The use of hierarchy in statistical confidentiality is null, and, we believe, that's unfortunate. Despite the fact that the research literature and current practice on disclosure limitation, as a whole, have identified many relevant aspects of the three fundamental objectives "maximize usefulness", "maximize safety" and "minimize cost", just few of those aspects are taken into account at the decision stage in the applications. For instance, a quick search in the statistical literature (Mackie and Bradburn 2000, Fienberg 2003 and 2004) reveals that a set of lower level objectives for "maximize usefulness" should contain at least:

- "Accessibility" : Who can access the data, under which condition (in terms of time, cost, technology, etc) ;
- "Feasibility": Data users ability to undo disclosure protection procedures for inferences about statistical models of interest;
- "Quality": Quality of data users inferences using the disseminated data as compared to the quality of data users inferences under the original data;
- "Transparency": The extent to which the data dissemination procedure provides direct or even implicit information on the added bias and variability induced by the procedure.

However how many applications do really refer to such hierarchy? But few exceptions, as Abowd and Woodcock (2001), the common approach is to focus on very few items of those mentioned above. "Accessibility", "Feasibility", and "transparency", for example, are often not considered and "Quality" is replaced by "Quality of parameter estimates" ignoring other relevant aspects of "Quality" already identified in the research literature like "Model Uncertainty" or "Quality of residual analysis", to mention just a few. Such incomplete hierarchies can compromise any posterior effort aimed to define suitable attributes and assessing sensible trade-offs.

We do dot believe on a universal hierarchy appropriate for any arbitrary disclosure limitation problem. In decision theory it is well understood, in fact, that even for a specific decision problem with a single decision maker, hierarchies are not unique and different hierarchies can lead to different courses of action.

However we does believe that statistical agencies that disseminate data collected under a pledge of confidentiality would obtain great benefits by building their own hierarchy for their specific data dissemination problems. The hierarchy would help the decision maker

to clarify the interpretation of the relevant objectives, to check that no relevant aspects of the decision have been ignored, facilitating the communication of all parts involved in the problem.

4 Defining and Selecting Attributes

Assuming that a set of objectives has been specified and that it is appropriate for the data dissemination problem of interest, the next step, in the decision analysis, is to define a suitable set of attributes. In this section we review the main issues related with attribute definition and attribute selection [2] and we discuss their relevance for data dissemination problems.

4.1 Types of Attributes

According to the research literature on decision theory, we can distinguish three types of attributes (see Keeney and Gregory 2005): *natural attributes*, *constructed attributes*, and *proxy attributes*.

When there exist an obvious scale that can be used to measure the extent to which an objective is achieved such a scale represents a *natural attribute*. For example, the objective "minimize cost" has the natural attribute "cost measured in euros". Natural attributes directly measure the extent to which an objective is achieved in a natural scale commonly understood. But for the example described above and few others, the use of natural attribute in disclosure limitation is quite unusual due to the complexity of the objectives involved in the problem. For example, for the objective "maximize usefulness", for which we have described a possible "splitting" in section 3, no natural attribute can be defined even after trying to decompose the objective in lower level objectives and searching for natural attributes for each of the lower level objective.

When no obvious scale for an objective exists, we could still try to directly measure the extent to which an objective is achieved by constructing a "subjective scale" or "subjective index". The scale, which is called *constructed attribute*, should take into account the relevant aspects of the objective as described by the hierarchy of the decision problem. A panel of experts usually take the responsibility for it. Two illustrative examples of constructed attributes - both discussed by Keeney and Gregory (2005)- are the Dow Jones Industrial Average that measures movement in the stock market, and the Michelin rating system for restaurants. Note that although defining a constructed attribute that directly measures the extent to which an objective is achieved is not always possible (or successful) "interpretability" is a priority for constructed attributes. "Interpretability" here means that the decision maker should be able to associate to each "level" of the constructed attribute a clear description of the consequences for the objective of interest and viceversa, the decision maker should be able to describe consequences for the objectives in terms of levels of the attribute (in the terminology used in subsection 4.2, "interpretability" requires the attribute to be *comprehensive* and *understandable*). To the extent of our knowledge there are not examples of constructed attributes in disclosure limitation and

[2] Our review of attribute definition and attribute selection is a very short summary of a more detailed discussion on the topic by Keeney and Gregory (see Keeney and Gregory 2005).

that's unfortunate as we explain in the next section. For the moment we turn our attention to proxy attributes.

A *proxy attribute* is an attribute that reflects the degree to which an associated objective is achieved but does not directly measure the objective (Keeney and Raiffa 1976). The value of a proxy matters only to the extent that it serves as predictor of the objective of interest. It's usefulness depend on the "prediction error" or, which is the same, on the relationship that exists between the objective of interest and the associated objective measured by the proxy and on the decision maker's understanding of such relationship. Note, for example, that *monotonicity* of the relationship (the greater/smaller the value of the attribute the better the achievement of the objective) is not sufficient. Monotonicity allows the decision maker to rank different alternatives in terms of the attribute but not to make sensible trade-offs. Trade-off assessment, in fact, requires to understand how differences in the proxy attribute values translate into different degrees of achievement of the objective of interest (for a formal argument in terms of utility functions se Keeney and Raiffa 1976, chapter 2).

The research literature on disclosure limitation presents several (we believe too many!) examples of proxy attributes. For example, for the objective "maximize usefulness" different authors have discussed proxy attributes based on (Hellinger, Kullback-Leibler and other) distances between a density estimation under the perturbed and unperturbed data (see Gomatam et al. 2003, for example). Others, for the same objective, have proposed proxy attributes based on measures of discrepancy between summary statistics for the perturbed and unperturbed data (see G. Crises 2004 for a review). The intuition underlying all these proxies is that low distortion of the original data implies approximately correct inferences for most of the statistical analysis. However it is hard to see how perturbations of the original data expressed by any of the proxies listed above can be translated by the decision maker into meaningful statements about degradation of relevant statistical inferences. Not even monotonicity is guaranteed to be preserved. Being this the case how can one responsibly think whether, for example, an increment of the Hellinger distance, say from 0.4 to 0.5, is worth an increase, say from 2% to 3%, in the percentage of records correctly re-identified? Wouldn't be better in this case trying to define a constructed attribute? We believe so and we explain why in section 4.3. To make the argument we need a preliminary description of the desirable properties of an attribute.

4.2 Desirable Properties of an Attribute

Keeney and Gregory (2005) identify five sufficient properties of good attributes. Because of space limitations, here we discuss just three of them: *comprehensiveness*, *understandability* and *operationality*.

An attribute is *comprehensive* if satisfies two properties: (a) it takes into account all the relevant aspects of the objective that is meant to measure; (b) the values judgments embedded in the attribute are appropriate for the decision problem. A constructed attribute that takes into account only differences in parameter estimates, for instance, is not comprehensive for the objective "Quality" according with the hierarchy outlined in section 3 since the users cost of the inference (which includes time to access the data, software and people skills necessary to analyze the disseminated data) are relevant aspects of

the objective not considered in the attribute. On the other hand, if for the same objective we use an attribute based on discrepancies between summary statistics evaluated using the perturbed and unperturbed data (as proposed in G. Crises (2004)) we are making the value judgement that data users will ignore the information provided about the masking and will use the released masked data sets *as if* they were the original data (otherwise the attribute should compare summary statistics under the original data with the corresponding *estimates* under the masked data). Assuming that the selected statistics considered in the attribute reflect the relevant aspects of "Quality" (we really doubt that such statistics do exists in real applications), the chosen attribute will be comprehensive to the extent to which this value judgment is appropriate for the decision problem. In general all the attributes that involve counting, such as "number of records re-identified", implicitly assume that all items are equally important and we should ask the question whether this is an appropriate assumption in the decision problem under study.

A comprehensive attribute takes into account all the relevant aspects of the corresponding objective but is not of much help in the decision analysis if all the parts involved in the decision problem do not have a clear understanding of the levels of the attribute. An attribute is *understandable* if the decision maker and anyone else interested in the decision process understands what each level of the attribute means in terms of the objective of interest. In Data Disclosure Limitation understandability is a key property for two reasons. First of all, if any of the attributes for the fundamental objective 'maximize safety" and "maximize usefulness" is not understood by the decision maker, then no sensible trade-off can be made. In addition understandability is a necessary condition to maintain data users' confidence on the agency's data dissemination procedures. This relates to the discussion on the "transparency" objective in section 3. Data users understanding of the perturbation that the data dissemination has introduced into the statistical analysis of interest is crucial for the acceptance of the statistical agency's data dissemination procedure.

Unfortunately, even an attribute that meets all the previous properties is not sufficiently good if it's practical use in the decision problem generates a cumbersome work for the analyst who is in charge to implement the decision analysis. A fundamental property of an attribute is thus operationality. An attribute is *operational* to the extent that it is possible to obtain the values of the attribute for the set of different alternatives. [3] For the fundamental objective "maximize safety", for example, a natural attributes for a Bayesian would be a function of the intruders' posterior distribution for the sensitive variables given the disseminated data. However it could be the case that evaluation of such posterior distribution is too cumbersome or even infeasible and constructed or proxy attributes should be considered instead.

[3]Note that in Keeney and Gregory (2005) the definition of "operational" attribute addresses the additional concern of whether the attribute allows the decision maker to make informed value trade-offs. The definition that we use here does not address this additional concern and rather refers to the definition of "'measurability" described in Keeney and Raiffa (1976). The reason for this choice will be apparent in the next section.

4.3 Selecting an Attribute: a Decision Problem

The evaluation of a candidate attribute in terms of the properties described in the previous section is not a dichotomic outcome (presence, absence). Different attributes are comprehensive (understandable, and operational) to different degree. In complex decision problems it is quite unusual to find attributes that fully satisfy all the properties. Rather the choice requires a decision about how much of a subset of properties we are ready to sacrifice to improve achievement of the others. A usual trade-off is the one that involves "understandability and comprehensiveness" on one hand, and "operationality" on the other. If a comprehensive and understandable attribute is not operational we might choose an alternative attribute which is not as much as understandable and comprehensive as the original but can be evaluated for the different alternatives. Note, however, that comprehensiveness and understandability are the priority. Meaning that we should reduce these properties as little as possible and stop as soon as we get an attribute that within the constraints of the problem (time, money, people skills, technology) is operational. The preference structure in the trade-off that we have just described, has a natural explanation in the discussion of the desirable properties in section 4.2. Comprehensiveness and understandability are a necessary condition for the decision maker to be able to make sensible trade-offs which is the core of the decision problem.

These ideas have a direct application on the prescriptive order in attribute selection. As described by Keeney and Gregory (2005), the choice of an attribute for a given objective should start with natural attributes. If, even after trying to decompose the objective in lower level objectives, no natural attributes can be found (or can be found but are not operational) then we should try to define a constructed attribute. Only when this turns out to be an infeasible task we should look for proxy attributes.

The discussion on section 4 shows that, having checked that no natural attributes can be defined, too often in disclosure limitation problems we choose the easiest "solution". We identify a proxy attribute. The non "interpretability" of proxies, and thus the practical impossibility to make sensible trade-offs do not seem a sufficient deterrent for their use nor a motivation to invest on constructed attributes. Part of the reason, we believe, is that quantitative proxies (such as those described in G. Crises (2004)) are perceived as "more objective" than subjective indices as constructed attributes are. The argument, however, seems weak. As commented before, knowing the value of some measures of discrepancy between distributions (or between summary statistics) evaluated for the perturbed and unperturbed data is, in general, of little or no value to understand the degradation of relevant statistical inferences. This is especially true for the complex statistical modeling of interest in the analysis of longitudinal linked data sets. We believe that in these cases, constructed attributes based on a panel of experts (that could certainly contain representatives of legitimate data users) would allow much more sensible trade-offs.

This is not meant to say that proxy attribute are useless. Rather than constructed attributes should receive more attention that they did so far.

5 Conclusions

On page 9 of the book, *Value-Focused Thinking. A Path to Creative Decision Making*, Keeney says:

> "There is a tendency in all problem solving to move quickly away from the ill-defined to the well-defined, from constraint-free thinking to constrained thinking. There is a need to feel, and perhaps even to measure, progress toward reaching a "solution" to a decision problem."

To get that feeling of progress, in Data Disclosure Limitation, we often quickly identify objectives, attributes and some viable alternative and proceed to evaluate them, without making the effort that a comprehensive definition of the problem, in terms of alternatives, objectives and attribute would require. This paper has addressed these concerns with particular emphasis on the importance of a proper structuring of objectives and the prescriptive order in the selection of attributes. The discussion hasn't focused on longitudinal linked data, as much as desired, but, we believe, it is particularly relevant for this type of data, given: (i) the complexity of the modeling usually associated to the analysis of longitudinal linked data; (ii) the multiple decision makers involved in the problem; and (iii) the different perspectives and perceptions of risk and utility that must be accommodated in the final decision.

Acknowledgements

Preparation of this paper was supported by the U.S. National Science Foundation under Grant EIA-0131884 to the National Institute of Statistical Sciences. The contents of the paper reflects the authors' personal opinion. The National Science Foundation is not responsible for any views or results presented.

References

Abowd, J. M. and Lane J. (2003a), "Synthetic Data and Confidentiality Protection", Workshop on Microdata, August 21-22 2003, Stockholm, Sweden.

Abowd, J. M. and Lane J. (2003b), "The Economics of Data Confidentiality". Unpublished paper presented at the National Research Council's Committee on National Statistics Workshop on Confidentiality and Access to Research Data Files, October 16-17, Washington DC.

Abowd, J. M. and Woodcock S. D. (2001), "Disclosure Limitation in Longitudinal Linked Data". In *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies.* P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (Eds.), North-Holland, Amsterdam, pp. 135–166.

Abowd, J. M. and Woodcock S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data". In *Privacy in Statistical Databases.* Domingo-Ferrer J. and Torra, V. (Eds.), Springer-Verlag, pp. 290-297.

Doyle, P. Lane J., Theeuwes J., and Zayatz L. V. (2001), *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*, North-Holland, Amsterdam.

Fienberg, S. E. (2003), "Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches". Workshop on Microdata, August 21-22 2003, Stockholm, Sweden.

Fienberg, S. E. (2004), "Datamining and Disclosure Limitation for Categorical Statistical Databases", IEEE International Conference on Data Mining, Workshop on Privacy and Confidentiality, November 1, Brighton, England.

Gomatam, S., Karr, A. F., and Sanil, A. (2004), "Data Swapping as a Decision Problem", *Journal of Official Statistics*, to appear.

G. Crises (2004), "Information Loss Measures for Microdata in Database Privacy Protection", Research Report CRIREP-04-004, Sep. 2004, Dept. of Computer Engineering and Mathematics, Rovira i Virgili University of Tarragona, Spain.

Keeney, R. L. (1992), *Value-Focused Thinking. A Path to Creative Decision Making*, Harward University Press 1992.

Keeney, R.L. and Raiffa H. (1976), *Decisions with Multiple Objectives*, New York:Wiley 1976.

Mackie, C. and Bradburn, N. (2000), *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* Committee on National Statistics (CNSTAT). Commission on Behavioral and Social Sciences and Education. Washington DC: National Academies Press, 2000.

Trottini, M. (2003), "Assessing Disclosure Risk and Data Utility: A Multiple Objective Decision Problem", UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg 7-9 April 2003.