ISSN 1023-098X
Cat. No. KS-CS-00-002-EN-N

RESEARCH IN OFFICIAL STATISTICS

2■2000

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

Printed in France

PRINTED ON WHITE CHLORINE-FREE PAPER

Research in Official Statistics

ROS — An international journal for research in official statistics

ROS — Volume 3 — Number 2 — 2000

Contents

Lead-lag estimation by means of the dynamic time warping technique

A. Varfis, L. Corleto, J.-M. Auger, D. Perrotta, M. Alvarez

Institute for Systems, Informatics and Safety,
European Commission. Joint Research Centre

Key words: dynamic time warping, lead-lag analysis, time series, leading indicators, forecasting

Abstract

An original approach to measure the lead-lag delay structure between two time series is described and analysed. The novel method is derived from the dynamic time warping (DTW) algorithm used in speech recognition. DTW warps the two series under study in order to achieve a better alignment of their ups and downs. Measuring the distortion undergone by the series provides a delay estimator. The proposed method is compared to an alternative delay estimation approach based on conventional tools such as linear transfer functions and correlation analysis. Two pairs of series have been investigated. One pair is made of synthetic data, whereas the other involves a real-life production index series and a leading indicator. The experiments run so far have featured several interesting properties that make DTW a promising approach for lead-lag structure analysis.

1. Introduction

1.1. Leading indicators

The leading indicator (LI) approach to economic and business forecasting consists in the identification and use of time series whose fluctuations anticipate those of the broad economic activity. Investigations in the field started in the early 1930s at the National Bureau of Economic Research, as a method to anticipate emerging states in the ongoing business cycles.

For a long time, focus in the use of LI was mostly put on the forecast of turning points. Indeed, advance warning signals of recession or recovery are undoubtedly of great value to managers and policy-makers. Making forecasts at all times is important, however, and nowadays LI are being used to produce predictions uniformly over time, and not solely to disclose turning points. Moreover, whereas the detection of turning points usually concerns forthcoming peaks or troughs — in which case the indicators are truly leading ones — ongoing or recently elapsed events may also be involved, whereby one speaks of coincident or lagging indicators, respectively.

Construction of indicators, or composite indexes thereof, is based on heuristics from economic theory. The expert in economy who builds the index selects variables which to some

extent are expected to lead the quantity of interest at hand. The latter will be referred to as main or target series (the vocable 'coincident indicator' is sometimes used in the econometric literature, but we deem it would be confusing here). A main series, typically, is some broad indicator of economic activity, as for instance production index or gross domestic product. No formal econometric model underlies the leading index or indicator construction, though. As mentioned in Newbold and Bos (1994), 'the leading indicator approach is sometimes despairingly referred to as measurement without theory'. It is thus conceivable that an indicator which purports to be a leading one eventually proves to belong to the coincident or lagging category, or fails to produce timely signals for some of the turning points (Klein and Moore, 1983).

In like manner, the lack of firm theoretical foundations entails several uncertainties in the relationship between indicator and target series, even when the former consistently anticipates the latter. To start, let us dispose briefly of an important discrepancy, which however is not the chief concern in this paper: between the two series, the relative size of matching peaks or troughs is neither stable through time, nor easily predictable. Our main topic then is about the remaining sources of inconsistencies, which relate to the lead-lag structure between the indicator and main series. Several issues are involved. First, the expected number of leading periods is not known a priori, not to mention that it is quite common to face delays for peaks and delays for troughs that are systematically different. Second, lead-lag values evolve markedly through time and the variability is hard to predict. Finally, measuring lag values may be difficult for most instances along the time series: relatively easy cases are turning points, which however do not correspond to sharp reversals, thereby making their precise localisation somehow ambiguous; hard cases occur over the much larger aggregate time span covering 'steady' periods of expansion or recession, where LI are also used for forecast tasks.

1.2. Background

The main object of this paper is to describe an original approach for lead-lag delay estimation between indicator and target series. The development of our novel methodology has its origins in a 'support to the Commission' research project for Eurostat where neural networks were the core modelling technique to be used for causality analysis and forecasting purposes (Varfis et al., 1998). The multi-layer perceptron family of connectionist models may be viewed as non-linear regression models, which in the present context would be fed with LI as regressors and target series as response variables. It will be convenient to use *X* and *Y* generic notations inspired from regression models to designate these series:

Y: The target series are short-term industrial production indexes. These are monthly series, each of which corresponds to one specific industrial branch and country.

X: The candidate LI series are issued from expectation surveys: Directorate-General for Economic and Financial Affairs from the European Commission runs business climate surveys where corporate managers are asked to provide ternary answers to questions like 'Order book level? (above average/average/below average)' or 'Expected Production

trend for the following months? (increasing/stable/decreasing)'. The balance between the percentage of positive and negative answers is then used to eventually produce quantitative *X* series from the qualitative survey results.

Associated series: It will be said that *X* and *Y* are associated in a given model whenever *X* is used as LI series for *Y*. Since this paper focuses on the delay estimation methodology, little attention will be paid to the exact nature of the series involved. We just mention here that for the Eurostat project associated series always had the same country and branch.

The causality analysis between associated *X* and *Y* series included several topics, as for instance investigating delays between turning points, or assessing the informative content of *X* for the purpose of forecasting *Y*. Building models for causality analysis would normally assume that the delay between associated *X* and *Y* series is stable, or varies slowly with time. Facing small samples — as happens here since macroeconomic associated series are involved — neither connectionist nor conventional approaches are able to cope with strongly fluctuating delays (or other forms of marked structural instability, for that matter). Loosely speaking, the reason is because machine learning models represent the input/output relationship via a 'transfer functional' of sorts from *X* towards *Y*. This is indeed a very general and powerful modelling paradigm, which however implicitly assumes a stable lead-lag structure, inasmuch as it is very unlikely and unnatural to expect delays to depend on mere *X* values. By and large, the variables — if any — or mechanisms that underlie the lead-lag changes are not in general very well understood or known, and the present task was no exception.

To get a first estimate of lead-lag values, we implemented an acknowledged traditional approach that consists in performing a correlation analysis via linear transfer function models (see Section 4.3). As measured by this method, delays between *X* and *Y* series exhibited a strong variability, both in terms of amplitude and frequency of changes. As a consequence, straight input/output learning models — either conventional or connectionist ones — were of little promise. This led us to look for alternative approaches that might have the ability to cope with the lead-lag instability issue.

1.3. Contents

The proposed novel methods to deal with the lead-lag structure of associated series are based on a technique called dynamic time warping (DTW). DTW essentially pertains to the field of speech recognition (Waibel and Lee, 1990), where it constitutes a mainstream technique against which connectionist methods have been benchmarked (Bottou et al., 1990). The use of DTW in other fields and different kinds of time series is definitely much more restricted and came to us via a recent paper from a collection of advances in data discovery and mining (Berndt and Clifford, 1996). The approach looked appealing enough to have it extended and adapted to the task at hand, as a means to tackle the lead-lag problem.

DTW is amply described in Section 2. In the frame of the Eurostat project (Varfis et al., 1998), DTW-based methods have been developed with two different goals. One objective

was to cope with strong lead-lag variations between leading and target series, or at least alleviate their negative consequences. Novel cross-fertilisation algorithms have been developed for that purpose, which mix DTW with conventional or connectionist clustering methods. These hybrid approaches, however, are not described here. In part because their exposition is lengthy and would make the paper extend markedly over the allowed space, and mostly since further investigations are needed.

Rather, we present in Section 3 another facet of DTW, which can be operated to measure lead or lag values. This original way to measure delays is benchmarked against an acknowledged traditional approach that consists in performing a correlation analysis via linear transfer function models. The novel tool for time series analysis is quite flexible, to the extent that it comprehends many tuning parameters.

The DTW-based and the conventional approach are tested and compared in Section 4.4, on true LI series as well as on synthetic data. Performance with actual series is admittedly more important than good behaviour with artificial data. Nevertheless, more attention is paid to the latter series. The reasons are detailed in Section 4.4, and essentially amount to the fact that lead-lag values are known only for the synthetic series.

2. Dynamic time warping

2.1. Genesis

In the seventies, researchers in the field of speech recognition have devised and used a dynamic programming approach that partially addresses difficulties resembling to some of the delay problems described in the introduction (Waibel and Lee, 1990). This approach is known as dynamic time warping, or DTW. The idea of applying DTW to macroeconomic time series started from a recent short paper by Berndt and Clifford (1996), where the wide scope of this new route for time series analysis was evoked. The main strength of DTW-based techniques lies in their ability to operate approximate pattern detection tasks, where imprecise matching is allowed along two ‘dimensions’. To give an early feeling of what is meant here, let us consider briefly a simple case where two patterns of the same size, $P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_n\}$, would be compared. The two dimensions along which approximate matching is possible are the following:

- **Amplitude**. This is part of almost every pattern detection technique. The matching criterion focuses on and allows for discrepancies between components with the same index. Most criteria take the form of $f(\sum_i d(p_i,q_i))$, i.e. an aggregate of component-wise dissimilarities. For instance, criteria based on the Euclidean distance belong to this category. For $i \neq j$, p_i and q_j are normally not compared.

- **Timing**. The chief peculiarity of DTW-based methods is the possibility to have distortions along the time — or longitudinal — dimension (that is along the i indices: DTW is essen-

tially employed with temporal patterns). So $d(p_i, q_j)$ terms with i (j may contribute to the eventual fit value. This results from the possibility to match one p_i component with several consecutive q_j components, or conversely. Some fuzziness is thus allowed in the matching process, inasmuch as sub-patterns of either series may 'wait' for appropriate subpatterns in the other.

By and large, the incentives for introducing DTW in speech recognition tasks came from the need to locate and match archetypal word templates within waveforms of actual speech. Different instances for the utterance of a given word may give rise to large timing changes for the speech signals, and recognition techniques had to be robust with that respect. Quite similar problems arise when characteristic patterns are searched in macroeconomic time series. For instance, real-life downturn or upturn points do not have a pure prototypical shape, say $\cap$ or $\cup$, nor anything very close to that. Human visual perception of turning point may embody a great deal of basic shapes and width, e.g. (Π or $\bigvee$, and is also robust to random fluctuation with relatively high signal-to-noise ratio. Yet, knowledge elicitation on numerical characteristics of turning points is not straightforward, and methods which are able to cope with time stretches and amplitude bursts are welcome.

2.2. Warping paths

Dynamic time warping is best explained with the graphical illustration of a warping path. First, let us consider two finite time series sequences that we wish to align:

$$S = \{s_1,\ldots, s_i,\ldots, s_{lS}\} \text{ and } T = \{t_1,\ldots, t_j,\ldots, t_{lT}\}$$

These sequences may be of different length (i.e. $l_S \neq l_T$ is possible) and do not play a symmetric role in the matching process. T is seen as a template series for which approximate instances have to be searched in S. The horizontal axis in Figure 2.1 carries the i indices of S elements, and the vertical axis represents the j indices of T elements. The thick grey line passing through points with (i,j) coordinates marks a $W = \{w_1,\ldots,w_k,\ldots,w_p\}$ *warping path*. A $w_k = (i(k),j(k))$ pair indicates that at the k^{th} step in the warping path $s_{i(k)}$ is aligned with $t_{j(k)}$. If the series have the same length, a perfect alignment would follow the main diagonal, which is represented by a black dashed line. The first few alignments exemplified in Figure 2.1 tell us that:

$k = 1$: t_1 is aligned with s_1. Actually this initial alignment is a quite common but not mandatory boundary condition (more on this shortly).

$k = 2$: t_2 is aligned with s_1. So S is stalled ...

$k = 3$: t_3 is aligned with s_1 ... for two steps (or three times).

$k = 4$: t_4 is aligned with s_2.

$k = 5$: t_4 is aligned with s_3. Now T is stalled.

So, every warping path corresponds to a sequence of index alignments. Given some $d(s_i,t_j)$ discrepancy measure between time series elements, the D degree of disagreement — or mismatch — between S and T along W is measured by the cumulative discrepancy along their warping path. The definition of d is influential and belongs to the hyper-parameters that the user may tune according to his needs. The absolute distance or its square root are good choices for the present purposes. If we use the latter choice, then D writes:

$$D(S,T,W) = \sum_{k=1}^{p} \sqrt{|s_{i(k)} - t_{j(k)}|} \qquad \text{(Equiation 2.1)}$$

The objective of a DTW procedure is to search a path that minimises Equation 2.1. Of course, some restrictions have to be enforced to delineate the space of admissible paths: for instance, trivial solutions are to be avoided, as for example the empty path or the one-step path corresponding to $\min_{\{i,j\}} d(s_i,t_j)$; most restrictions, however, are common sense conditions which shrink the set of allowable moves that the dynamic programming algorithm has to explore at each step and thus avoid combinatorial explosion in the path search. The gist of the search algorithm is given at the end of this section. Before, we rather focus on the set of constraints or boundary conditions used in most of our experiments, which is listed and commented below. In Figure 2.1, full black lines mark either frontier paths or some of the forbidden moves.

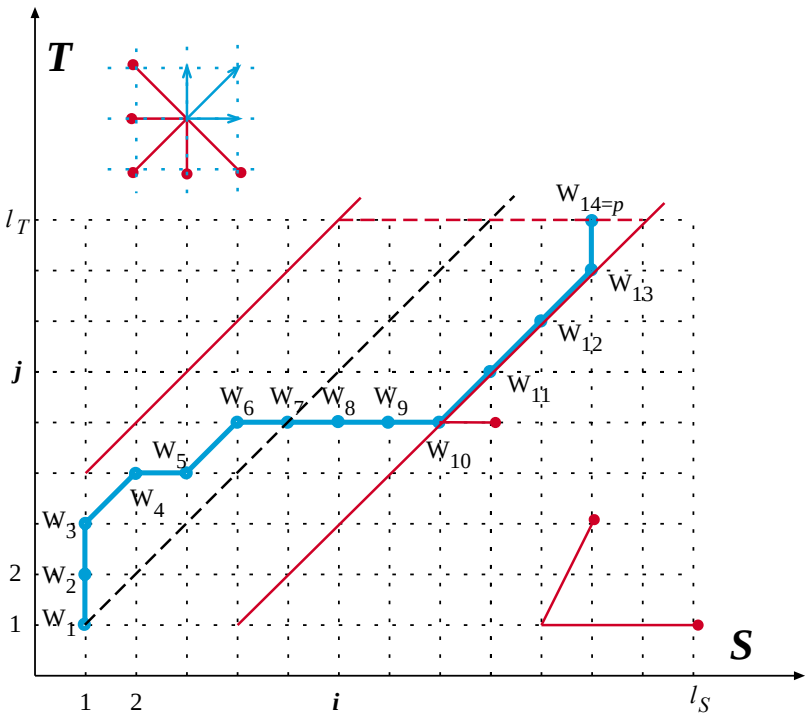


Figure 2.1: Illustrative warping path

1. **Monotonicity**. The warping path should correspond for both time series to an increasing sequence of index alignments: $\forall k \in \{2,p\}$, we require $i(k) \leq i(k\text{-}1)$ and $j(k) \geq j(k\text{-}1)$. Either time series may get stalled but backward steps are not allowed. This is illustrated in the upper-left sketch in Figure 2.1 where admissible one-step moves are represented by grey arrows whereas forbidden one-step moves are depicted with black segments ended by a dot.

2. **Continuity.** Only one-step moves are allowed: $i(k)\text{-}i(k\text{-}1) \leq 1$ and $j(k)\text{-}j(k\text{-}1) \leq 1$. The bottom-right sketch in Figure 2.1 displays two monotonically increasing yet forbidden steps. Only the three steps depicted in the upper-left sketch fit both the monotonicity and continuity requirements.

3. **Template spanning.** $j(1) = 1$ and $j(p) = l_T$. Together with the continuity requirement, these boundary conditions entail that every element of T is used at least once in the warping path. Full use of T is not so much a constraint but rather pertains to the template concept, whereby it is sensible that a search for T instances in S naturally precludes the template pattern to be excised. Conversely, either or both tails of S may eventually not belong to the warping path. In our illustrative example, the warping path ends as $j(p) = l_T$ — the l_T ordinate limit is marked with a black horizontal dashed line — whereas $i(p) < l_S$.

4. **Anchor points.** To implement the dynamic programming algorithm that searches for the shortest path it is necessary to preset a starting alignment point. In the illustrative example of Figure 2.1 we use the most common choice for that matter, which consists of constraining $w_1 = (1,1)$ — i.e. t_1 and s_1 are aligned. Note that additional anchor points may possibly be enforced according to user's requirements. The algorithmic starting alignments point, however, raises several specific issues.

   - The basic DTW algorithm which operates with a starting left-hand anchor point can be used to implement other variants. For instance, a right-hand start solution can be computed from the basic algorithm by reversing both series.

   - There is normally no reason why the user should know where on S the approximate template matching starts. A common task indeed is to identify such starting points when instances of a pre-stored template are searched within a long S time series. The search procedure would then run the DTW algorithm sequentially for each candidate starting point on S. Therefore, in the (default) left-hand anchor setting, s_1 refers to the starting point of the working sub-series, whether or not it comes from a larger series that is being repeatedly censored on its left.

   - The diagonal of perfect timing fit — a black dashed line in Figure 2.1 — starts at w_1. It corresponds to the $i = j$ line with respect to the local i coordinates of S alluded to in the previous bullet. In like manner, the warping window concept that is described below is also relative to the current left anchor point.

5. **Warping window**. To avoid excessive timing discrepancies and to reduce the search space, it is common to set constraints on the $i(k)$-$j(k)$ differences. Symmetric limits are often used, e.g. $|i(k)-j(k)| \leq 3$ as depicted by the black limiting diagonals in Figure 2.1. At w_{10}, the T template could not keep on being stalled as indicated by the forbidden black horizontal segment. In the sequel, the ω symbol will be used to indicate a strictly positive symmetric warping window size ($\omega = 0$ does not permit of warping and corresponds to conventional pattern matching. So sheer DTW with a warping window restriction implicitly assumes $\omega \geq 1$). In the illustrative example, we have $\omega = 3$.

To conclude the description of DTW, which essentially is dynamic programming applied to the matching of temporal patterns, let us present briefly its search algorithm. To ease the exposition, we shall specialise on the constraints and boundary conditions described in the above list.

Each (i,j) coordinate on the $l_S \times l_T$ dashed grid in Figure 2.1 may be related to two values. In the first place, we attach to (i,j) the $d(s_i,t_j)$ discrepancy measure. Second, we may compute recursively $cum(i,j)$, which we define as the lowest cumulative discrepancy amongst admissible partial paths starting at the w_1 anchor point and ending with a (s_i,t_j) alignment. Notice that D values as in Equation 2.1 correspond to complete paths, that is $cum(i,j)$ instances for which $j = l_T$, according to the template spanning condition. In view of the monotonicity and continuity constraints, it is easy to recognise that cum obeys to the following recurrence relation:

$$cum(i,j) = d(s_i,t_j) + \min(cum(i\text{-}1,j\text{-}1), cum(i\text{-}1,j), cum(i,j\text{-}1))$$

Initialising with $cum(1,1) = d(s_1,t_1)$, which reflects the starting anchor point condition, the grid may be filled progressively from left to right and bottom to top. Outside the warping window limiting diagonals, cum values need not be computed.

The searched minimum value for D, say $D_{\min}(S,T)$, is the lowest from the $7=2\omega+3$ cumulative discrepancies for the grid points in the upper dashed line, where $j = l_T$ and $|i-l_T| \leq 3$. The corresponding coordinate gives us the right tail of the $W_{\min}$ optimal warping path. We may then build backwards the complete $W_{\min}$ path, by tracing recursively the admissible previous point that has the lowest cum value.

3. Lead-lag measurements

Whereas a warping path display as in Figure 2.1 is appropriate for describing the mechanisms of DTW, it hardly helps us to visualise the S and T series and their alignments. In this section, we provide alternative means to represent the outcome of a DTW process, which at the same time illustrate and describe how time series of delay estimates are produced. Our explanations will proceed via a guided capture of Figure 3.1.

Let *S* and *T* be the patterns to be aligned by means of DTW. In the frame of a study on leading indicators, typical instances of *S* and *T* could be intervals extracted from associated *Y* and *X* time series, respectively. The upper plot of Figure 3.1 represents a synthetic pair of such 15-dimensional patterns ($l_T = l_S = 15$). Visually speaking, the dashed line (*T*) and the full one (*S*) have similar shapes towards their ends, that is roughly along their first four and their last four components. On the other hand, the series behaviour in the middle part is not in phase. Conventional pattern matching methods, which typically use fit criteria that are based on cumulated component-wise errors, would yield for this illustrative example error terms that are small for the bordering indices and large for the central ones. Alternatively, some traditional methods could admit the dashed series to be shifted — here two position towards the right is appropriate, which would improve the fit in the middle part at the expense of the error terms on the edges. Conventional approaches involve rigid patterns, which impedes a better alignment of the series peaks and slumps throughout the whole time spell. Conversely, DTW methods allow for moderate changes in regime during the matching periods, to the extent that either series pattern may be warped in order to reach a better alignment.

The series of the upper plot underwent a DTW matching procedure with the default specifications detailed in Section 2.2 and $\omega = 2$ as warping window size. Let *i, j* and *k* denote again the running indices of *S, T* and *W*, respectively. In the second plot of Figure 3.1, both warped series are plotted against the warping path index.

We have coined development plot such a display of $s_{i(k)}$ and $t_{j(k)}$ against *k*. Indeed, warping is implemented in DTW by enabling either series to be stalled during the matching process, i.e. to stay at its current value for more than one time step and subsequently resume from there. Stated differently, when one series is stalled, it may be seen and understood as if it were stretched (developed) through time, inasmuch as the stalled value is drawn out to match several values of the companion series (conversely, the companion series may be perceived as if it were squeezed).

Let us interpret now the development plot from the template series perspective (dashed line). Both possible perspectives give of course rise to dual interpretations, although the DTW algorithm is not symmetric in *S* and *T*.

From $k = 2$, the template remains stalled for two more steps: its value $t_{j(2)} = t_2 \approx -0.7$ is matched not only to the second component of the *S* series — nothing unusual so far — but also to the third and fourth components of *S*. The sequence of three constant values at ≈ -0.7 is depicted by a small horizontal dotted segment, which stretches the visual representation of the dashed line. In like manner, small horizontal dotted segments that occur later on within the full line correspond to stall periods for *S*.

The main objective of a development plot is to picture the match of peaks and troughs after completion of a DTW alignment. This does not require both series to share the same *y*-scale. Rather, different offsets are preferably used on the *y*-scale in order to segregate the series lines and enhance the visual perception.

At this stage, a little reasoning reveals that one stalling step for either series amounts to gaining one unit of lead with respect to the companion series. This is how DTW produces estimates of lead-lag values. The third plot of Figure 3.1 displays the algebraic lead of *T* over *S*, which is the line of (*i*(*k*)-*j*(*k*)) against *k* for the current DTW procedure. One observes, for instance, that the initial template lead of two time units remains stable over seven steps. This lasting stability is readily visible from the oscillating yet parallel evolution of both series during that period, in the development plot, which in turn discloses the same parallel behaviour in the corresponding part of the upper plot, if we were to shift the dashed line by two time units towards the future.

The second and third plots share the same *x*-scale, which is based on the warping path index. It is certainly more convenient, however, to visualise the time series of delay estimates on the natural time scale, to parallel the series displays in the upper plot. This is simply done and illustrated in the bottom plot, where stalling periods for the template — which correspond to augmenting lead — are represented by vertical segments, as against diagonal increasing segments previously. In the sequel, we shall use the latter representation to display the line plots of delay estimates. The representation is not symmetric in *S* and *T*, and picture the lead of the template series over the companion series, which also translates as the lead of the LI over the target series when *S* comes from *Y* and *T* from *X*.

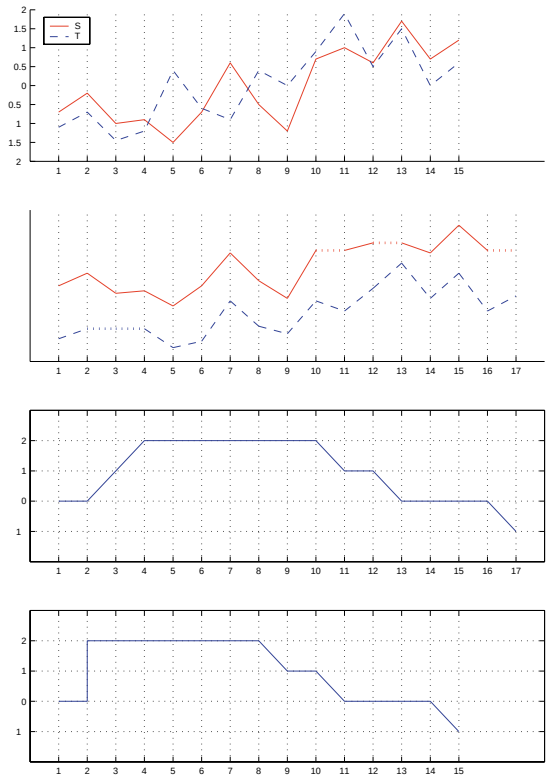


Figure 3.1: Development plot

4. Experiments

4.1. Several issues

Let us consider two associated series of interest, say *Y* and *X*. For the task of lead-lag estimation, the *S* and *T* sub-series to be matched by DTW will span the largest common range of *Y* and *X*.

By and large, DTW can perfectly be run with a short template series, as illustrated in the previous sections (*S* may be much longer than *T*, but not really shorter since $l_S \geq l_T - \omega$ is required for a solution to exist. The latter inequality is the sole — and very weak — condition for our version of DTW to be operable). Short templates are indeed employed in pattern detection uses of DTW where occurrences of a prototypical *T* pattern are searched along a large *S* series. The chief care to take then consists in selecting a reasonably low figure for the warping window size, lest the patterns become excessively distorted in regard to their length.

For delay analysis, we advocate alignments over a long range, which has two advantages of sorts. The first — minor — benefit is that the archetype speech recognition tasks which led to the design of DTW and experienced several success stories with it indeed operated template matching methods over large patterns. The second and more palpable advantage is that long-range alignments enable us to select larger warping window sizes without running the risk of achieving distorted and meaningless alignments due to data idiosyncrasies on short series. As a consequence, the method has — potentially — the ability to handle large lead-lag delays, which is a clear-cut possibility when macroeconomic leading indicators and target series are involved.

For any pair of series with the same length, a DTW process will always issue a time series of delay estimates. Merely getting a solution is thus not informative as for its quality, which should be assessed. Two aspects are involved: primarily, the value of our delay estimates; indirectly, the quality of the DTW matching fit, which presumably affects the delay estimates as well.

Regarding the latter point, quantitative criteria exist (see, for example, Berndt and Clifford, 1996). Yet, the issue for lead-lag analysis is rather to assess qualitatively whether the DTW alignment is sensible. This cannot be achieved in any circumstances, and even when a good fit is potentially feasible, taking appropriate pre-processing steps is often crucial. Indeed, carrying out a full match of two series that exhibit markedly different shapes is rather pointless, especially when alignments over long patterns are planned. Before running a DTW procedure over full series ranges, both series should undergo appropriate transforms — if any — to achieve a common general appearance. This issue is briefly illustrated in Section 4.2, when the real-life series are described.

Assuming that a sensible DTW process has eventually been run, it remains to assess the delay estimates, which is a harder task. For one thing, the actual lead-lag values are unknown

for real-life data. Furthermore, the concept of delay amongst a leading and a target indicator series is somehow ambiguous, especially during periods when the business activity is in 'steady state'. As a consequence, absolute estimator accuracy can only be measured on synthetic pairs of series. Relative performance with respect to alternative methods is also difficult to figure out. Principled approaches to estimate delays exist, but no one is even close to the status of a reference methodology, or is acknowledged as the technique to benchmark against. It is therefore questionable to compare, on real data, our DTW-based delay estimations to more conventional ones. Again, resorting to artificial data is hard to avoid.

In the following subsections, we briefly describe in turn: the specific series that have been used to test our novel delay estimation technique, one alternative approach based on traditional methods, and the outputs of either technique for the pairs of series at hand.

4.2. Series

Two pairs of series are investigated in the text. One pair is made of synthetic data, whereas the other corresponds to real-life data from the Eurostat project, picked amongst the many pairs of associated series that have been analysed. We shall rather focus on the former set. Admittedly, a litmus test should preferably be run on real examples, which typically are more complex and representative than artificial ones. As evoked previously, however, it is extremely difficult to assess a novel method on associated series whose actual delays are not known, especially when there is some evidence that these delay series are quite volatile. The synthetic series are thus used to have the true lead-lag structure under control and assess properly the methodologies involved. Conversely, the real-life delay series are essentially plotted for illustrative purposes. Hereafter follows a short description of both pairs of series.

4.2.1. Artificial series

X and Y both have 280 elements that are generated as follows (for the artificial series, we simply have $Y=S$ and $X=T$):

X consists of 280 i.i.d. samples from the uniform distribution between – 5 and 0.5: $X(t) \sim Unif_{(-0.5,\,0.5)}$.

To build Y, we first partition the time in three periods, namely $P_1 = \{1,\ldots,160\}$, $P_2 = \{161,\ldots,200\}$ and $P_3 = \{201,\ldots,280\}$. Each period is further subdivided into four quarters (of consecutive elements), for example for the first period, we define $P_{1,1} = \{1,\ldots,40\}$, $P_{1,2} = \{41,\ldots,80\}$, etc. Then Y is generated according to the following pseudo-code:

1. for $p = 1$ to 3
   - if $t \in P_{p,1}$, then $Y(t) = X(t + 2)$ (i.e. X lags Y by two time steps)
   - if $t \in P_{p,2}$, then $Y(t) = X(t)$
   - if $t \in P_{p,3}$, then $Y(t) = X(t – 2)$

- if $t \in P_{p,4}$, then $Y(t) = X(t + 1)$ (and $Y(280) \sim Unif_{(-0.5, 0.5)}$, which stands for the unobserved $X(281)$ value.

2. We add a random noise to Y: $Y(t)$ ($Y(t)$+ ((t) where $\varepsilon(t) \sim Unif_{(-0.15, 0.15)}$.

So the algebraic lead of X over Y cycles three times through the same sequence of four delay values, namely {-2, 0, 2, –1}. The idea behind these cycling sequences is to investigate the 'frequency of change' factor, which is lowest during P_1 and highest along P_2. The full series of actual delay values is depicted by a wide background light grey line in Figure 4.1.

4.2.2. Real-life series

Indicators refer to intermediate goods industry for Germany. The production index (Y) is compared against the balance between positive and negative answers for the 'expected production' question given in Section 1.2 (X). The z-score transforms of these two series are displayed in the upper plot of Figure 4.2. The plot below represents both series after an effort to remove seasonality and align their cycles by means of the following transforms:

$$S = \nabla_{12} \log Y = -\log \frac{Y_{t-12}}{Y_t} \approx \frac{Y_t - Y_{t-12}}{Y_t}$$

$$T = SA(X)$$

The transform on Y is issued from two loose pieces of domain knowledge. First, a question on 'expected trend' refers to flow quantities — as against level ones — and thus involves ∇Y differences of the production index. Second, as elicited in Dossé and Maquet (1995), there is some evidence that corporate managers who answer to trend questions tend to refer implicitly to the same period one year before.

According to the above heuristics, the raw X series and the transformed S series should exhibit approximately paralleling patterns (as they eventually do in the middle plot of Figure 4.2). However, the X candidate template series exhibits a seasonal component, which definitely has to be removed before proceeding. For one thing, it would be odd to match on the long range a series that has kept its seasonal component (X) with a series whose seasonal component has essentially been cancelled out by the ∇_{12} operator (log Y). Furthermore, were S to possess a marked seasonal component as well, it would still be sensible to remove seasonality from both series involved, lest the alignment be mostly driven by the seasonal peaks and troughs.

Summing up the previous considerations, it is important in the present experimental setting to remove the seasonality component from X, yet without affecting the phase of X. *SA* stands for such a seasonal adjustment method. It is a seasonal index method based on estimates of monthly averages after removal of the trend by means of a centred 13-point moving average.

Visual inspection of the second plot of Figure 4.2 suggests that both series are not strongly out of phase and do not exhibit any marked seasonality: running DTW is now sensible.

4.3. Conventional approach

The alternative delay estimation approach to benchmark against DTW is based on conventional tools such as linear transfer functions and correlation analysis. In the remainder, our bespoke conventional method will be referred to as CORR. Its description will be very succinct, since these concepts are old and widely documented (see, for example: Box et al., 1994; Kendall and Ord, 1990; or Pankratz, 1991). The prewhitening stage has essentially been implemented as in Ljung (1995). Starting from the *S* and *T* transforms of the original *Y* and *X* series, and denoting by $\{t_1,\ldots,t_l\}$ their largest common range (200-odd elements), we carry out the following operations:

Prewhitening: First, an autoregressive model (AR) of order 15 is fitted to *T*: being *B* the backshift operator, a ϕ polynomial of order 15 is searched which makes $U = \phi(B)T$ as white as possible. The same polynomial is applied to *S*, yielding $V = \phi(B)S$. Let us make four short comments on this prewhitening stage: series that are not properly filtered give rise quite easily to spurious cross-correlation coefficients, as is known since the paper of Box and Newbold (1971) which criticised the methodology used in Coen et al. (1969); when many series have to be analysed, the need to automate the identification process motivates here the use of large AR models, as against more parsimonious ARMA prewhitening filters; an AR order larger than 12 has been selected to possibly cope with residual seasonality effects, i.e. not removed by the transforms leading to *S* and *T* from *Y* and *X*; the filter is still applied for the leftmost elements, with some reweighting, so that the length of the new *U* and *V* series is still *l*.

Windowing: Let us denote by $U_{W(t)}$ the sub-series (or window) of *U* containing 19 consecutive elements centred at time *t*: $U_{W(t)} \div \{U_{t-9},\ldots,U_t,\ldots,U_{t+9}\}$. Similarly for $V_{W(t)}$. If such a window is shifted by steps of size one across the series, the set of time indices for possible $(U_{W(t)}, V_{W(t)})$ pairs of sub-series becomes $\{t_{10},\ldots,t_{l-9}\}$. For each of these pairs:

Impulse response: We compute the impulse response function for positive and negative lags up to eight time steps. Notice that these $\{\beta_{W(t),k}\}_{-8\le k\le 8}$ impulse response values are proportional to the cross-correlation coefficients.

Delay estimate: The value of *k* that corresponds to the largest absolute value of $\beta_{W(t),k}$ is taken as delay estimate for time *t*: $delay(t) = \text{argmax}_{-8\le k\le 8} |\beta_{W(t),k}|$.

The range of the above delay estimation model is {-8,+8}. The DTW estimator was set to have the same range by using $\omega = 8$ as warping window size.

4.4. Results

4.4.1. Artificial series

Figure 4.1 displays the time series of delay estimates produced by DTW and CORR for one pair of synthetic series. The wide background light grey line indicates the true sequence of delay values. The thin black line corresponds to the CORR estimator. The thicker grey line represents the DTW estimations, which are represented as explained earlier for the bottom plot of Figure 3.1: increasing delays are displayed as vertical segments, whereas decreasing delay values correspond to steep diagonal segments.

A great deal of pairs of artificial series have been generated by varying the state of the random number generator, and the behaviour of their delay estimators has been inspected. The illustrative example was selected because it concentrates the following recurrent features observed during the session:

1. Both methods faithfully perform well over P_1 (for the notation see Section 4.2.1) along which actual delay values remain constant over 40 time steps. DTW consistently outperforms CORR over P_2, which corresponds to the highest frequency of delay change in our setting. The conventional approach regularly fails there, as is plain for the illustrative example, whereas DTW often manages to catch rather closely the right sequence of delays. As for the intermediate P_3 regime, both methods usually cope well with changes every 20 time units, but DTW still experiences less failures than CORR.

2. Even when a regime change is correctly identified, both methods often make small errors in either direction as for the precise time of delay shift. For instance, DTW slightly leads and CORR markedly lags the jump from $P_{1,2}$ to $P_{1,3}$ at step 80 (whose precise position is located at the centre of the wide grey line). Conversely, CORR leads and DTW lags the ensuing change of regime at step 120.

3. CORR often exhibits sharp bursts close to a change of regime, as at index 220 in the illustrative example (and also, to a lesser extent, at time 40).

4. Small peaks frequently occur along the DTW curve, as for instance twice during $P_{1,3}$ as well as in the rightmost segment.

5. Due to the windowing operation (see Section 4.3) CORR cannot issue forecasts at either end of the series: for our windows of size 19, estimates are not available for the first and last 9 time indexes. Conversely, DTW series of delay estimates span the full range of the presented series.

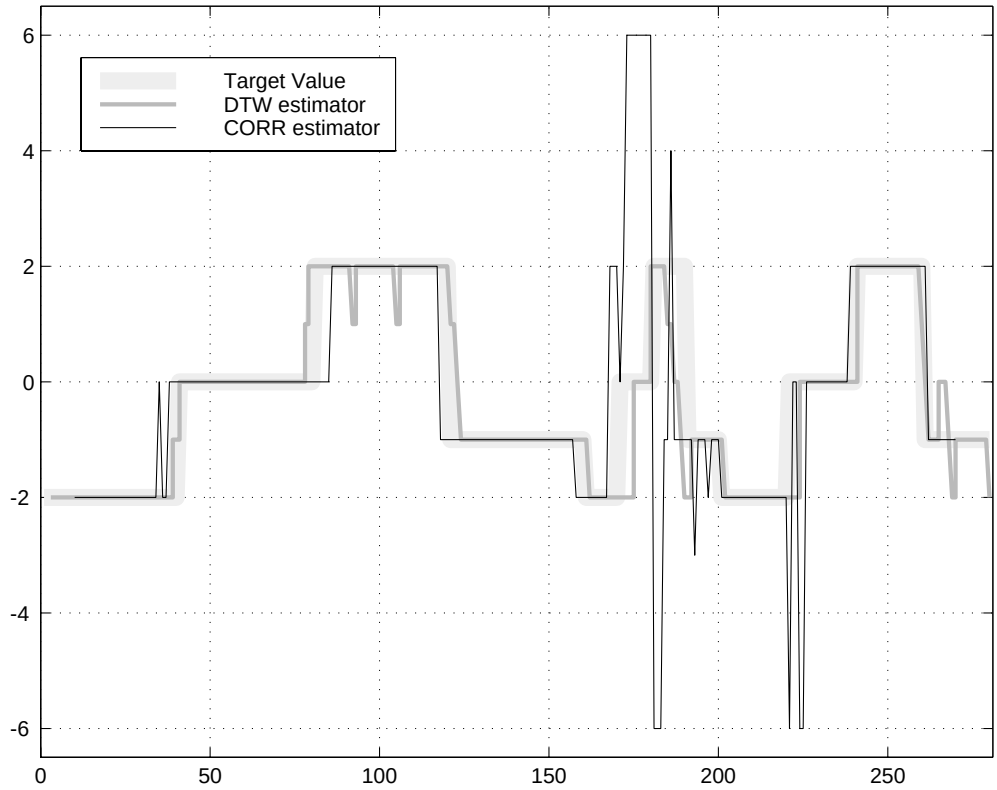


Figure 4.1: Comparing DTW and CORR on artificial data

Regarding the first point, the behaviour of CORR with respect to the pace of change is easy to interpret in qualitative terms. The conventional approach operates by computing impulse response terms over a window of both series, which understandably gives unpredictable results when the window overlaps many different delay regimes. One could think of using narrower windows, but this approach soon fails due to sample size requirements for a correct estimation of cross-correlation coefficients from noisy data. Indeed, experiments with tighter windows showed that spells of erratic delay estimates are likely to occur, even during the 'steady' P_1 period.

DTW proved to be much more flexible. To come closer to limits, further tests were run in which the shorter cycle had length 20. More often than not, DTW managed to track the sequence of delays during the period when changes occur every five time steps. As expected, CORR performed very badly there. On the other hand, the ability of DTW to adapt quickly also results sometimes in data idiosyncrasies being fitted, as exemplified by the small peaks evoked at point 4. This pertains to the stability versus plasticity dilemma (or trade-off). We are presently investigating elaborate versions of the algorithm where a penalty term for stalling steps is added to the DTW payoff function. This would smooth out most annoying small bumps, yet possibly at the expense of the adaptation potential.

In addition to plasticity, another characteristic that differentiates CORR from DTW is the 'continuity' of sorts owned by time series of delay estimates produced by the latter approach. The continuity condition of Section 2.2 (point 2) ensures that indices are never skipped along the warping path. Considering also the companion monotonicity condition, it follows that $(i(k)\text{-}j(k))$ delay estimates may vary by only (1 unit when a stalling step occurs. DTW is capable of tracking gaps or jumps in the lead-lag structure by letting the same series stall for several consecutive steps. In particular, abrupt increases of the algebraic lead of *T* over *S* look like being fit instantly, merely because our choice was to depict rising leads by vertical segments when we plot delay estimates on the natural time scale. Yet, modelling abrupt changes has a cost. The peculiar approximate 'continuity' we are trying to describe here is induced by the price to pay for modelling jump shifts, which in turn results from the following considerations: every warping path step adds one *d* discrepancy value to the objective *D* cumulative distance (see Equation 2.1); a change of (units in the $(i(k)\text{-}j(k))$ difference corresponds to (stalling steps (for the same series); by and large, taking non-diagonal steps tend to lengthen the final warping path, hence increasing the *D* aggregate.

DTW thus returns time series of delay estimates that underwent a kind of continuity constraint. Continuity is a priori a valuable property, at least insofar as one would reasonably expect true sequences of lead-lag values to behave rather smoothly. Actual settings where delay values change frequently yet without jump shifts are certainly much more favourable to DTW than to CORR. Conversely, if the underlying true lead-lag structure involves very abrupt delay shifts, say from – 5 to + 6 (so still within the (– 8, + 8) reachable range), it is conceivable that the DTW algorithm may need something more to find benefits in charging 11 idle *d* terms to D_{min}. The typical stimulus would be the true delay sequence staying close to + 6 for some time, to let DTW compensate the previous investment with a series of low payoff terms. In like manner but the other way round, it will be hard for the estimator to come from – 5 if preceding true delay values are not large lags as well. To summarise in broad terms the previous heuristic considerations that have been tested only partially, we may state that DTW is expected to fare much better with step-like delay patterns, as against peak shapes (the latter being anyway a very unnatural lead-lag structure). We saw previously that the same holds for CORR, yet in a more demanding form since constant values over larger plateaux are essentially needed. The only advantage we see for CORR in modelling complex lead-lag structure is its insensitivity to the jump magnitude. But are large delay shifts realistic?

To conclude this section, let us consider point 2. It is insightful to interpret the small timing errors made by either approach in the vicinity of changepoints. These estimation shifts are due to data idiosyncrasies, which may give rise a different phenomenon depending on the method. Regarding CORR, windows that have their centre close to a changepoint correspond to an impulse response function dominated by two peaks, one for each lag involved. With noisy data, it is then plausible that the larger peak — which indicates the delay estimate — will not always correspond to the tiny majority of instances where, without noise, $Y(t) = X(t\text{-}true\_delay)$ would hold exactly. Since CORR estimates are not subject to any 'continuity' condition, these confusions do not necessarily take the form of a mere step timing error. The peak at step 36 illustrates an isolated mistake. Beyond the above comments which describe

how the CORR estimator is confused, we do not know at this stage how to explain the magnitude of much larger bursts that occur frequently at changepoints, as for instance around step 220.

Regarding DTW, the development plot (Figure 3.1) illustrates how the algorithm is driven by the series oscillations, which are approximately set in phase when the current delay is estimated correctly. To minimise the cumulative payoff, an alignment that is lost because of a regime change has to be recovered by DTW, via non-diagonal warping path steps. Noise in the data may generate spurious ups and downs that let DTW go slightly astray and anticipate or postpone the required stalling steps. Notice that contrary to CORR, DTW will hardly generate isolated peaks that are due to delay shifts, especially when the jump or gap exceeds one unit: aggregating discrepancy terms along the two-way path soon becomes too expensive in terms of cumulative payoff to be worth the detour. Isolated peaks of magnitude as evoked at point 4 are more frequent. These are also due to spurious fluctuations, yet can occur everywhere along the time series of delay estimates.

4.4.2 Real series

The salient features of DTW and its pros and cons with respect to CORR have just been amply discussed. On the other hand, as forewarned at end of Section 4.1 and start of Section 4.2, applying our methods on real series is hardly conclusive at the present state of our knowledge. The real-life example will thus be dealt with in a succinct manner and essentially for illustrative purposes.

The upper and middle plots of Figure 4.2 display the original X and Y series, and their T and S transforms, respectively. A description of the original series and rationales for their transforms have been discussed in Section 4.2.2.

The bottom plot represents the DTW and CORR time series of delay estimates. The lack of coherence amongst the estimators is striking. Yet there is a priori no reason to consider one method to be better than its contender. Nor should we take as granted that either approach is really suitable for the data at hand, assuming that a model for the lead-lag structure exists.

5. Conclusion

A new idea for estimating the lead-lag structure between time series has been described and investigated in this paper. The novel approach is based on fresh principles inspired from the dynamic time warping (DTW) algorithm, commonly used in speech recognition. Even though the scope of the algorithm is in no way restricted to speech recognition tasks, few investigations have been performed outside of the latter field. Our interest in DTW has been raised by Berndt and Clifford (1996), who point to the potential of DTW to detect patterns within generic time series. Our use of DTW is different, since we require both series to have similar overall patterns, possibly after appropriate transforms, and match them along their

full — or a long — range. The resulting DTW alignments then provide an original way to measure the lead-lag structure between the series under study.

As for every innovative technique, a great deal of testing is required to assess properly its advantages and drawbacks, and to mature and enhance the original ideas. As a consequence, neither can firm conclusions be drawn at this stage, nor is the novel algorithm close to its ultimate version. Still, the experiments run so far have featured several interesting properties that make DTW a very promising approach for lead-lag structure analysis. Furthermore, ongoing studies suggest that many features of the algorithm are prone to undergo effective modifications to enhance the estimator properties.

In this paper, we describe an inaugural assessment exercise that has been carried out with artificial data, over which a rather unsophisticated version of DTW was benchmarked against a method based on correlation analysis (CORR). Even though prototypical characteristics of real-life delay sequences are not known, it seemed reasonably realistic to us to generate synthetic data according to the following specifications: lead-lag values would fluctuate through time, lest the task be trivial; delay values would shift at various rates, including rather high frequencies; the magnitude of any single shift would be moderate; additive noise is present in the observations. By and large, DTW outperforms CORR in the previous setting, particularly so during periods of higher rate of delay change. Indeed, CORR locally require some steadiness in the delay sequence, to estimate correctly the coefficients of the underlying linear transfer function model. Conversely, DTW is not based upon a local parametric model, and demonstrates to have a good plasticity. Another potential advantage of the previous 'non-parametric' feature of DTW lies with its ability to generate estimates at the series ends, as against stopping short by a half window when symmetric windows are used in CORR. Having recent right-hand estimates is certainly valuable for forecasting purposes. The performance of DTW in the latter perspective has not been tested so far.

DTW produces time series of delay estimates that exhibit a continuity of sorts. Since DTW is a dynamic programming method, a precise formulation for the property is hard to elicit. The continuity attribute is rather appealing, not only because estimated delay curves are smoother and their trend is easier to visualise, but also because a 'continuous' truth seems to be more realistic. Conversely, CORR sometimes produces quite erratic estimation sequences, since there are no constraints on the difference between successive delay estimates. This latter feature might be helpful, were the truth to involve large delay shifts. We plan to investigate the robustness of DTW against abrupt delay changes, even though we deem that wide jumps are hardly realistic. The few experiments run so far suggest that CORR might sometimes fare better than DTW, at least when changes occur at some befitting intermediate frequency (large frequencies penalise CORR; low ones eventually enable DTW to model any delay shift within its reach, which depends on the warping window size).

DTW for lead-lag analysis is an intriguing approach that requires additional testing along two different perspectives. On the one hand, the experimental setting to benchmark the method or assess its performance should be extended and refined. One could, for instance, generate artificially many different ARIMA time series with increasing coloured noise and

with different lead-lags, and compare methods on this basis. Or one could search for real time series where DTW and CORR give similar results, add disturbances, and observe performance degradation. The second perspective is about improving the method itself. Here investigations have started on several issues. As we just did for one of these in the above section, we present now very briefly two topics under study, with the gist of the first findings.

The definition of the *d* discrepancy measure is rather influential on the smoothness of the DTW estimator. Underweighting the influence of larger mismatches produces smoother curves. By and large, the square root of the absolute distance (see Equation 2.1) is a good choice.

We are developing the methodology to generate elaborate versions of the algorithm where additional terms — not depending on s_i or t_j — are appended to *d*. For instance, knowing of the proper manner to include penalty terms for stallings steps in the payoff function enables the user to get control on the plasticity-stability trade-off.

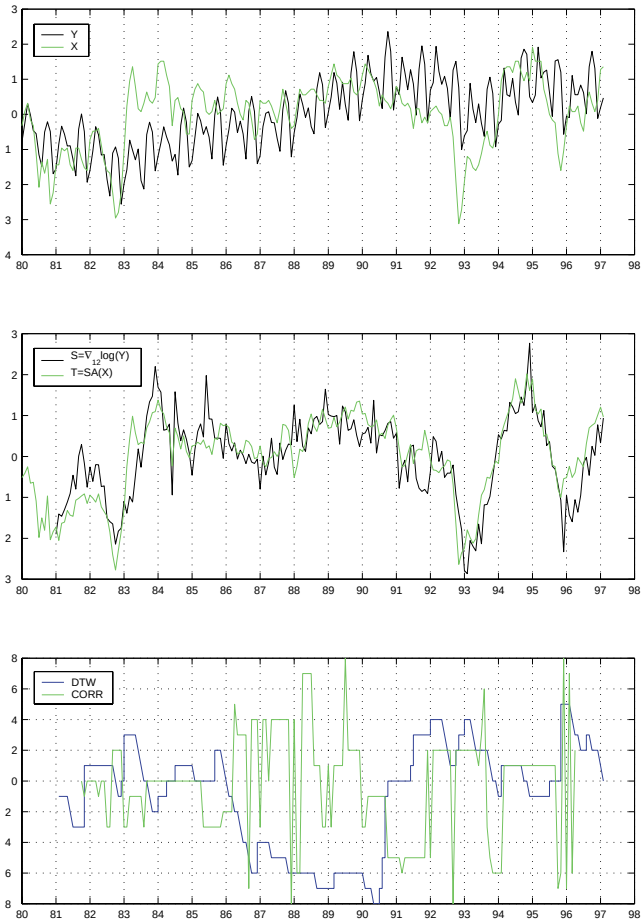


Figure 4.2: Comparing DTW and CORR on monthly indicator series. x-axis marks are years.

6. References

Berndt, D. and Clifford, J.(1996), 'Finding patterns in time series: A dynamic programming approach', in Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, Chapter 9.

Bottou, L., Fogelman Soulié, F., Blanchet, P. and Liénard, J. (1990), 'Speaker-independent isolated digit recognition: Multilayer perceptrons vs dynamic time warping', *Neural Networks*, 3, pp. 453–465.

Box, G., Jenkins, G. and Reinsel, G. (1994), *Time series analysis. Forecasting and control*, Prentice Hall, third edition.

Box, G. and Newbold, P. (1971), 'Some comments on a paper of Coen, Gomme and Kendall', *J. R. Statistical Society* A, 134, pp. 229–240.

Coen, P., Gomme, E. and Kendall, M. (1969), 'Lagged relationships in economic forecasting', *J. R. Statistical Society* A, 132, pp. 133–163.

Dossé, J. and Maquet, I. (1995), 'Rapport sur l'exploitation des données qualitatives de la DGII', Technical report, Eurostat — D2(ISTI).

Kendall, M. and Ord, J. (1990), *Time series*, Edward Arnold, third edition.

Klein, P. and Moore, G. (1983), 'The leading indicator approach to economic forecasting — retrospect and prospect', *Journal of Forecasting*, 2, pp. 119–135.

Ljung, L. (1995), *MATLAB*(, *System Identification Toolbox*.

Newbold, P. and Bos, T. (1994), *Introductory business and economic forecasting*, South-Western Publishing.

Pankratz, A.(1991), *Forecasting with dynamic regression models*, Wiley.

Varfis, A., Corleto, L., Auger, J.-M., Alvarez, M. and Perrota, D. (1998), 'Causality analysis with neural networks', technical report, call for tender N. 96/S 99-57617/EN, lot N.17.

Waibel, A. and Lee, K. (eds) (1990), *Readings in speech recognition*, Morgan Kaufmann. (Includes six papers about DTW).

Data and metadata transformations*

H. Papageorgiou, M. Vardaki,

Department of Mathematics, University of Athens, 15784 Athens, Greece
E-mail: hpapageo@cc.uoa.gr; mvardaki@cc.uoa.gr

Fragkiskos Pentaris

Department of Informatics, University of Athens, 15771 Athens, Greece
E-mail: frank@di.uoa.gr

Key words: statistical metadata modelling, metainformation systems, automated data retrieval, data quality, Internet

Abstract

The advantages of collecting structured metadata are presented and various ways of storing metainformation are discussed. An example of a statistical metadata model is given using the unified modelling language (UML) and properties and semantics of this model are further examined. The simultaneous manipulation of both data and metadata is discussed by introducing a set of transformations, including the addition and selection of data, the addition and removal of a variable and the grouping transformation. We argue on the benefits of using metadata transformations and, as a case study, we show how these transformations can be used in building a metadata-aware Internet web site that will support ad hoc retrieval of statistical aggregates.
The results of this paper were obtained within the framework of Eurostat's projects 'integrated documentation and retrieval environment for statistical aggregates (Idaresa)' and 'integrated public information systems and statistical services (IPIS)'.

1. Introduction

Metadata and metainformation are two terms widely used in many different sciences and contexts. In all those cases, these terms are defined as data about data (Grossmann, 1999). That is, metadata are every piece of information needed for someone to understand the meaning of data. The term was actually used in statistics for many years (Ghosh, 1988), but personnel working in large statistical offices only recently started to realise the potential benefits of formally producing, storing and using statistical metadata. Currently, all major statistical institutions are trying to increase the use of metadata by their employees. It is now acknowledged that metadata not only may affect the quality of the produced statistical indices (Papageorgiou et al., 2000b), but are important features in every statistical information system that aims in producing ad hoc statistical tables (Eurostat, 1993; Froeschl, 1997; Pa-

* The results of this paper were obtained within the framework of Eurostat's projects "Integrated Documentation and Retrieval Environment for Statistical Aggregates (Idaresa)" and "Integrated Public Information Systems and Statistical Services (IPIS)".

pageorgiou et al., 2000a). Consequently, metadata are related with the ever-increasing needs of data consumers to obtain more information of better and asserted quality. Furthermore, they play an important role in offices intending to provide new type of on-line services such as Internet-enabled access to selected parts of their aggregated data.

The automation of the processing of macro-, meso- or even microdata is based on the capability of designing a system that can manipulate both data and metadata using a predefined closed set of operators (transformations). These transformations simultaneously operate on both data and metadata, thus effectively producing new data with a known meaning (metadata). Research in data and metadata transformations is in early stages, which is partially due to the fact that designing and implementing a transformations-enabled system requires good knowledge of both statistics and informatics.

In this paper, we provide the reader with an overview of the structure of a metadata model and how we use metainformation, together with metadata transformations to automate the construction of statistical tables. For the sake of simplicity, we avoid getting into the mathematical details and instead, as a case study, we designate the way transformations are used while building an Internet web site of a statistical office. Initially, we give a description of how metainformation can be captured in a structured way and argue on the advantages of keeping metainformation structured. In the next two sections, we give an example of a metadata model and designate a set of data and metadata transformations that can be used together with this model. In the last section, we consider the case of building an Internet site for presenting ad hoc statistical aggregates, in order to demonstrate the usefulness of the presented transformations.

2. Types of metadata

Metadata are every piece of information needed for someone to properly understand the meaning of data. It is obvious that without metadata, data are of little value, since no one will be able to properly use them. However, there seems to be a mismatch between the needs of data producers and data consumers. The first ones, struggle to reduce the burden of their work and therefore, hesitate to fully document with the appropriate metadata the tables they produce. After all, this information is already kept in their minds. On the other hand, data consumers need data of high quality. This does include the metainformation needed to properly understand the produced data. If some pieces of this information are missing, then the consumers will have to guess the missing metadata; something that certainly leads in data of reduced quality (Papageorgiou et al., 2000b).

To solve this logical mismatch, statistical offices used to document their data using verbal-text notes and tables' footnotes. However this approach was inadequate since, quite often, verbal text proved ambiguous. Furthermore, it was not always clear how these footnotes would be affected during any processing of data. Finally, the lack of standards led to incomplete documentation of the produced data (missing metadata). That is, quite often, the data consumers were unable to find the exact piece of metainformation needed, as this information was originally misplaced or even not captured.

2.1. Structured metadata

To solve the previously mentioned deficiencies, Sundgren (1991, 1996) proposed the use of metadata templates. This was the first true attempt to capture metadata in a structured way. The advantage of this approach was the reduced chances of having ambiguous metadata as each field of the templates was well documented. Furthermore, the proposed templates were meant to be filled-in, each time a survey or data-processing procedure was carried out. This fact reduced the problem of missing metadata and helped in increasing the awareness of metadata among the statistical offices' personnel. Nevertheless, in recent years, researchers in the area of statistical metadata are proposing a new improvement: the use of statistical data and metadata models to capture, store and process metainformation.

Templates succeed in capturing, in a structured way, metadata. However, they have limited semantic power, as they cannot natively express the semantic links between the various pieces of metainformation. To capture the semantics of metainformation, a metadata model must be used. In this case, metainformation is modelled as a set of entities, each having a set of attributes. For example, for an aggregated table, these entities could be the sampled population, the appropriate variables and measurement units, etc. The real advantage comes from the fact that these entities are interrelated. This enables the user to follow a navigation style browsing, in addition to the traditionally used, label-based search. Furthermore, as we will later show, it is the core idea that enables metadata to be used in automated (i.e. human-free) data-processing procedures.

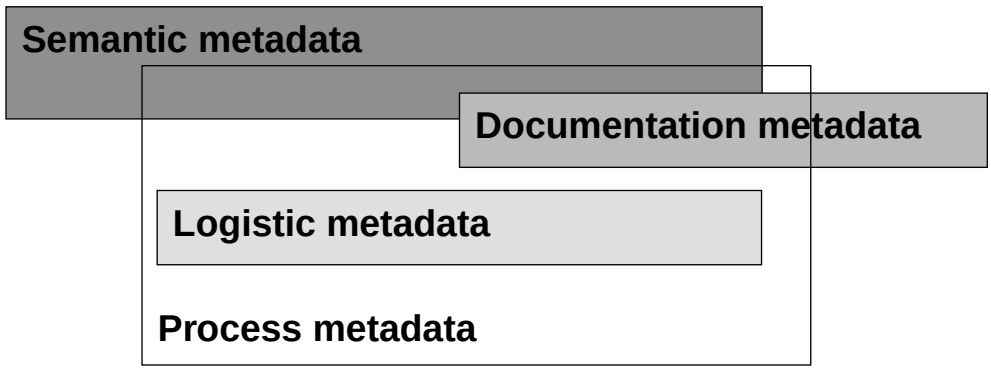
Currently, there is no widely accepted 'standard' for designing a statistical data model. This is a major drawback as it severely reduces the inter-operability of the proposed models and reduces the compatibility of the existing and forthcoming statistical information systems. In any case, it seems that the necessary metainformation is divided into the following four overlapping categories (Figure 1) (see also Kent and Schuerhoff, 1997):

Semantic metadata These are the metadata that give the meaning of the data. Examples of semantic metadata are the sampling population used, the variables measured, the nomenclatures used, etc.

Documentation metadata This is mainly text-based metainformation like, for example, labels that are used in the presentation of the data. Documentation metadata are useful for creating user-friendly interfaces, since semantic metadata are usually too complex to be presented to the user. Usually, an overlap between the semantic and documentation metadata occurs since, many times, we have to store metadata in both structured (i.e. semantic metadata used mainly by machines) and verbal-text (i.e. documentation metadata used by humans) form.

Logistic metadata These are miscellaneous metadata used for manipulating the data sets. Examples of logistic metadata are the data's URL, the type of RDBMS used, the format and version of the used files, etc.

Process metadata Process metadata are the metadata used by information systems to support metadata-guided statistical processing. These metadata are transparent to the data consumer and are used in data and metadata transformations. In the rest of the paper we will focus our attention on this type of metadata.



2.2. Modelling metadata

While building a metadata model, we have to decide on three major issues. That is, (i) what kind of modelling methodology we should use, (ii) how we are going to store the captured metadata and of course, (iii) what information is worth capturing (Layzell and Loucopoulos, 1989). The first two topics are related with Information Technology (IT) whereas the latter is related with statistics.

As far as the first question is concerned, there are two modelling methodologies that are currently used: the entity-relationship (E-R) model and most recently, the object-oriented model.

The E-R model has been widely applied and is part of many contemporary information systems. The basic concepts of this model are those of entity, attribute and relationship. An entity is defined as 'something about which information is recorded' (Layzell and Loucopoulos, 1989). Every entity has properties that are expressed in terms of attribute-value pairs whereas relationships are the associations between two or more entities. The E-R model offers the least semantics expressiveness power, but produces models that are simple and easily stored in relational database management systems (R-DBMS) such as Oracle®, Microsoft SQL Server® and Informix® (Date, 1990).

The adoption of object-oriented (OO) programming languages and object-relational DBMSs has reduced the popularity of the aged E-R model, which is now starting being replaced by new proprietary models (Stonebraker, 1994). The latter offer enhanced semantic richness, follow the OO modelling concept and provide for non-relational data stores. The experience of the authors with the Idaresa project (Eurostat, 1997a), which used an enhanced E-R mod-

el for storing metainformation, shows that using an E-R model for modelling metainformation is not an optimum choice, as it lacks the flexibility needed by a model to be adapted in unforeseen needs of the consumers for additional metadata. Therefore, it seems that currently, the best solution for building a metadata model is to use a designing tool which follows the OO model, like for example the UML (OMG, 1999).

The second question of how we are going to store metadata has two possible solutions. The first one is to use a DBMS such as Oracle® or ObjectStore® to persistently store our data. The advantage of this approach is that modern DBMSs offer high levels of robustness and low total cost of ownership. However, most DBMSs lack some important features, mainly they do not support full-text search engines and provide no support for metadata/schema versioning. In a recent SUP-COM project (Eurostat, 1998), the authors used a relational DBMS to store part of the metadata included in Eurostat's 'New Cronos' database (Eurostat, 1997b). Although the RDBMS and hardware used were rather fast, the end result indicated that it was not possible to build a single metadata database capable of holding all the metadata in New Cronos. The problems encountered were mainly related with the complexity (large number of table joins) of the SQL queries used. Consequently, it seems that a hybrid solution, where part of the metadata will be held in a DBMS and the rest will be stored using proprietary methods, is better. An example of such proprietary method is the use of extendable mark-up language (XML) and its derivatives. Although the technologies needed for using XML are still evolving, this approach seems quite promising. XML is already used for storing metadata in many different cases such as medicine and digital libraries (W3C, 2000). Consequently, using XML to store statistical metadata is becoming more and more appealing to system designers that seek a unified way of storing metainformation.

Finally, the last but most difficult question that must be answered while building a metadata model is what metainformation is worth capturing. If we oversize our model, we will end up with capturing information that is (almost) never used, thus effectively making a useless increase of our personnel workload. On the other hand, if we undersize our model, then the end-users will face a missing metadata problem which will severely reduce the usefulness and quality of the captured metadata (Papageorgiou et al., 2000b). However, it is extremely difficult to predict the needs of the users. For example, there are cases where a user will think that the provided metadata are sufficient for his work whereas a second one will believe that the same data cannot be used due to lack of documentation!

The IDARESA and IMIM (Eurostat, 1997a) projects both produced a model with hundreds of relationships and/or entities. Nevertheless, this volume of metainformation was still insufficient to solve the problem of documenting every possible statistical table. This fact clearly designates the difficulty of choosing what metainformation is worth capturing.

3. An example of a metadata model

Consider the data given in Table 1, which is an abstract of the macrodata available for government research and development (R&D) appropriations by Eurostat (1995, 1996). These

data are collected from the Member States of EU. For each country, R&D appropriations are classified according to year, type of appropriation (final or provisional), amount (in million ECU or equivalently in million EUR) and socioeconomic objectives according to the NABS'93 nomenclature (Eurostat, 1994). This nomenclature is a three-level hierarchical classification of objectives. The first level (chapter's level) classifies them into 13 chapters, which are further subdivided into a number of sub-chapters, and sub-sub-chapters.

A proper documentation of this table would take a model with hundreds of classes and relationships (Karge, 1998; Froeschl, 1997). However, to illustrate the main features of a metadata model, we use the one presented in Figure 2. This figure shows a description of a source frame using UML. A source frame represents a table containing data and metadata. Each source frame contains data for a sampling population. This population is a set of statistical units that was created using a sampling technique from a given statistical population, which is the one that we are really interested in. In the previously mentioned example, the statistical population is the 15 Member States of the EU. Since the number of the Member States is not significant, we usually do a survey and therefore, the sampling and the statistical population are the same. However, a sample can be easily selected using a set of criteria (sampling technique), thus defining a different sampling population.

Table 1: R&D data (Eurostat, 1995, pp 172–177. Eurostat, 1996, pp. 222–227).

Country	Year	Type of appropriations	Socioeconomic objective (NABS chapter)	Amount in million ECU
B	1994	final	8	32
DK	1993	provisional	8	55
DK	1993	provisional	9	944
GR	1993	provisional	6	21
E	1994	final	10	642
F	1994	final	7	882
IRL	1993	provisional	6	14
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
IRL	1994	final	1	49
NL	1994	final	3	115
A	1994	final	3	34
P	1994	provisional	1	24
FIN	1994	provisional	7	237
S	1994	provisional	2	120
UK	1994	final	5	72

If the elements of the sampling population are considered as the rows of the source frame, then the columns will be the conceptual variables. In the previous example, conceptual variables are the characteristics of the R&D appropriations, i.e. the socioeconomic objectives, the type of appropriation, etc. (see Table 1). We require that a source frame should contain at least three conceptual variables (specifying the temporal axis, the spatial axis and one further conceptual variable) that all have the same domain (sampling population).

Each conceptual variable is measured at a specific granularity level that is designated by its grouping level. In our R&D example, the grouping level of the conceptual variable (column) used for designating the socioeconomic objective is the chapters of the NABS nomenclature. In a different example, we could have used the NABS sub-chapter level, thus producing data of the finest granularity level.

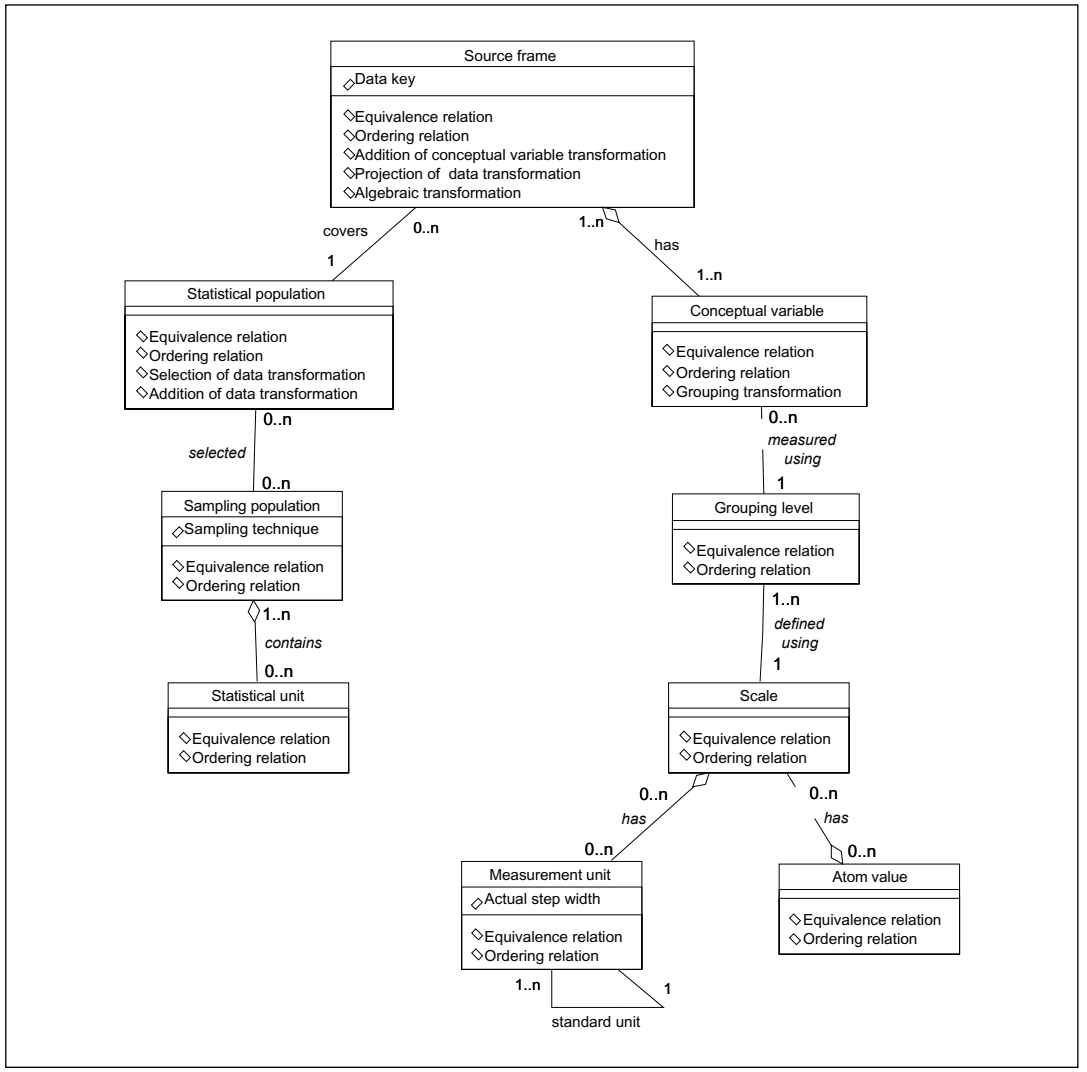


Figure 2. An example of a metadata model.

The set of all possible values of a measurement for a specific conceptual variable is a scale. In our example, the scale for specifying the socioeconomic objective is a set of 13 levels, i.e. 'Chapter 1, Exploration and exploitation of the earth, Chapter 2, Infrastructure and general planning of land-use, …, Chapter 13, Defence'. Each of these labels is an atom value. However, most of the times, a scale will be a subset of the real numbers. For example, the scale for measuring the amounts (see Table 1) is a numeric one. In this case, the elements of the scale are measurement units. These units have one important attribute, the actual step width, which designates the precision of the measurement (for example million ecu). Furthermore, each measurement unit might be related with a second measurement unit, which defines the standard unit, i.e. a different unit that is the preferred one for measuring the same characteristic.

4. Metadata operations

In Figure 1, we have used the notation of UML to give some of the methods (operations) that the previously mentioned metadata classes support. For example, the source frame class supports the equivalence relation, the ordering relation, the addition of conceptual variable transformation, the projection of data transformation and algebraic transformations. On a typical model, each class will support several methods. However, in our model, we illustrate only a small fraction that includes the most important ones. The definition of every transformation has four parts, i.e. a pre-condition, semantics, function definition and a post-condition:

Pre-condition This is a condition that must be satisfied for the transformation to be valid. The condition is designated with the help of the defined relations, like for example the ordering or the equivalence relation.

Semantics This is a documentation of what are the semantics of the transformation. For example, the semantics of the addition of data transformation is that the table is extended to cover a larger statistical population.

Function definition This is the definition of the data and metadata manipulation/processing that is required for the physical implementation of the semantics of the transformation.

Post-condition This is a condition that must be satisfied upon completion of the transformation. The post-conditions are useful in verifying the correctness of the outputs of a transformation as well as in optimising chains of transformations (see Section 6).

4.1. Equivalence and ordering relations

In Figure 1, we include two basic relations. These are the equivalence relation and the ordering relation. The first one designates whether two pieces of metainformation are equivalent. For example, the euro and million euro measurement units are equivalent if the first unit is measured with an accuracy (actual step width) of EUR 1 000 000 and the second unit is mea-

sured with an accuracy of EUR 1 000 000. The equivalence relation is extremely important and should be defined for every class of the metadata model. If a class of the model has no such relation, then there will be no automated way of testing for statistical data equivalence, since the equivalence relation of a class is defined in turns of having equivalent subparts. For example, two source frames (tables) are equivalent if they have equivalent statistical populations and equivalent conceptual variables respectively. Furthermore, two statistical populations are equivalent if they have equivalent sampling populations, etc. Consequently, if the metadata model has a class with no equivalence relation, there will be no way of cascading the equivalence relation from the top (source frame) to every sub-part of it.

The second relation, namely the ordering relation, designates when a piece of metainformation is 'better/preferable' than a second one. For example, a measurement unit with a thinner actual step width (precision) is better than the same measurement unit with a more granular actual step width. A different example is that the NABS nomenclature at the sub-chapters level is preferable than the NABS nomenclature at the chapters level. The ordering relation may not always be applicable. For instance, it is not accepted by everyone that the first version of the NABS is better than the second one. Nevertheless, the ordering relations (when they exist) are a useful tool for defining the pre- and post-conditions of transformations.

4.2. Data and metadata transformations

To automate the processing of statistical data, a set of operators must be defined. Examples of such attempts that operated mainly on the domain of data are given in Malvestuto (1993), Ozsoyoglu (1989) and Ghosh (1986). The difference to data and metadata transformations is that the latter are manipulations that affect both the data and metadata of a dataset. The transformations represent elementary processing steps that can be done on statistical data tables. However, because they designate what happens to both data and metadata, they can be nested to describe more complex data processing activities. In Figure 2, we refer to six transformations, namely the addition of a conceptual variable, the addition of data, the selection of data, the projection of data, the grouping transformations and algebraic transformations.

4.2.1. Addition of conceptual variable and projection of data

The addition of a conceptual variable to an existing source frame is a simple transformation where a dataset is extended with an additional column. Adding a new conceptual variable is only possible if the Statistical population of the source frame has such a measurable characteristic. If the values of the new variable were not measured, then the values of the new column will be equal to a special atom value, namely 'value not measured (missing values)'. A special case exists if the added conceptual variable is dependent (Date, 1990,p. 530) on other existing conceptual variables. It is then possible to derive the missing values from existing data (if available). For example, it is possible to add a new conceptual variable in the public financing of R&D source frame called 'Percentage of annual change in the R&D allocations'. The values of this conceptual variable can be mathematically calculated using existing data.

The term projection of data has its origins in the relational databases model. In our case, projection is a transformation that removes a conceptual variable, i.e. it is the opposite of addi-

tion of conceptual variable transformation. Typically, this transformation is applied on a source frame to remove the conceptual variables that are part of the data key, whenever individuals disclosure is required.

4.2.2. Addition and selection of data

Many times, the user has two source frames with equivalent conceptual variables but different statistical populations and wants to create a new source frame that will contain the data from both frames. Under certain circumstances, it is possible to append the data of the first table to those of the second table. The resulting source frame has a new statistical population that is the union of the two statistical populations of the original frames. The prerequisites of applying this transformation are difficult to be satisfied if the statistical populations are not disjoint. However, even in this case it might be possible to merge the two datasets, if the source frames have a data key, i.e. a combination of conceptual variables that hold sufficient information to distinguish duplicate tuples.

Semantically, the selection of data transformation is the opposite of addition of data. Using this transformation, someone can reduce the statistical population of a source frame, effectively decreasing the number of rows of the dataset. In this case, the resulting source frame has a new statistical population with the statistical units not satisfying, the selection criterion, removed.

4.2.3. Grouping transformations

The grouping transformations are operations that change the grouping level of one or more conceptual variables. There are two types of grouping transformations. The first type alters conceptual variables that do not belong to the data key of the source frame. During such a procedure, a mapping/conversion table designates how the previous values will be replaced by the new ones. If the new grouping level is better or equal to the existing one (see Section 4.1), there is no information loss. However, if this is not the case, the mapping will be imperfect and several previously measured values will be lost.

The second type of grouping transformations alters conceptual variables that belong to the data key of the source frame. In this case, after the values are converted, there might be tuples with the same data key. For example, assume that we have data collected at NABS subchapter level and we convert the data to NABS chapter level. In this case, every value measured for Subchapter 1.x. will be mapped to Chapter 1. Consequently, after the conversion is completed, there will be many different values for Chapter 1, each representing a different subchapter of Chapter 1. Therefore, whenever a grouping transformation affects a conceptual variable that belongs to the data key of the source frame, a second merging step is required (see also Lenz and Shoshani, 1997). Although, there are many cases where this second step cannot be calculated, often a simple function such as sum, average or minimum will be sufficient.

4.2.4. Algebraic transformations

The last type of transformations mentioned in Figure 2, is the algebraic transformations. By this term, we mean all the remaining transformations that can be applied onto a source frame.

A typical example of an algebraic transformation is the multiplication of the values of a conceptual variable by a constant deflator. The number of algebraic transformations is unlimited, and therefore human intervention is most of the time required to assert the correctness and meaning (metadata) of these transformations.

5. Benefits of using data and metadata transformations

Data and metadata transformations are the foundations for automating the processing of statistical data by machines. In a typically large statistical office, data consumers may have difficulties to select the appropriate data from several tables that look alike and have only minor, yet important, differences. However, in a metadata-aware system, the data consumer will have to use a kind of metadata query language, or even better, a suitable graphical user interface, to search for suitable tables, which will in turn reduce the dangers and ambiguities of verbal text. Furthermore, if the user produces a new table by using the data of one or more existing tables, a transformations-enabled system will automatically produce most of the documentation of the new table, thus further reducing the dangers of human errors.

A different application of data and metadata transformations is in automatically producing a needed table from a set of existing ones. In this case, the user describes the new table (i.e. gives the metadata of the new table) via the help of a declarative language or graphical user interface and the system finds all the possible ways of constructing the new table from the existing ones. This is a major improvement over the traditional way of specifying via a procedural language how the new table is created. Additionally, the system may use an optimiser to select an almost optimum way of producing the new table, taking into consideration the resources needed to perform each transformation (i.e. cost of retrieval of local or remote data, CPU processing needs, available network bandwidth, temporary storage needs, etc.).

The benefits of the above applications demonstrate the usefulness of transformations in answering the following two questions:

(i) Are we retrieving the right data?
(ii) Are we retrieving the data right?

This, in turn, allows for an enhancement of services' quality that the statistical offices offer to their data consumers.

6. A case study

As a case study, we describe the way transformations and metadata can be used in building the web site of a large statistical office. The use of the web allows for minimising the latency between the time when data are ready for the public and the time when decision-makers have acquired them. This is an important feature for statistics and therefore, all large offices are working hard building web sites for their data (Wouter et al., 1999). However, further de-

velopments are needed before the NSIs' web sites are ready to cover the demands of the consumers (Lamb, 1998). One of the deficiencies of existing web sites is that the data they offer to the users are static: few sites support data aggregations and data selections. However, none offers the chances of producing really new tables, by combining information retrieved from the existing ones. This is because information is stored in plain, yet huge, relational tables. Instead of using a metadata model, the metadata are held as plain labels and are used only for the sake of building nice user interfaces. However, a transformations-aware site can go further and allow for the creation of ad hoc tables.

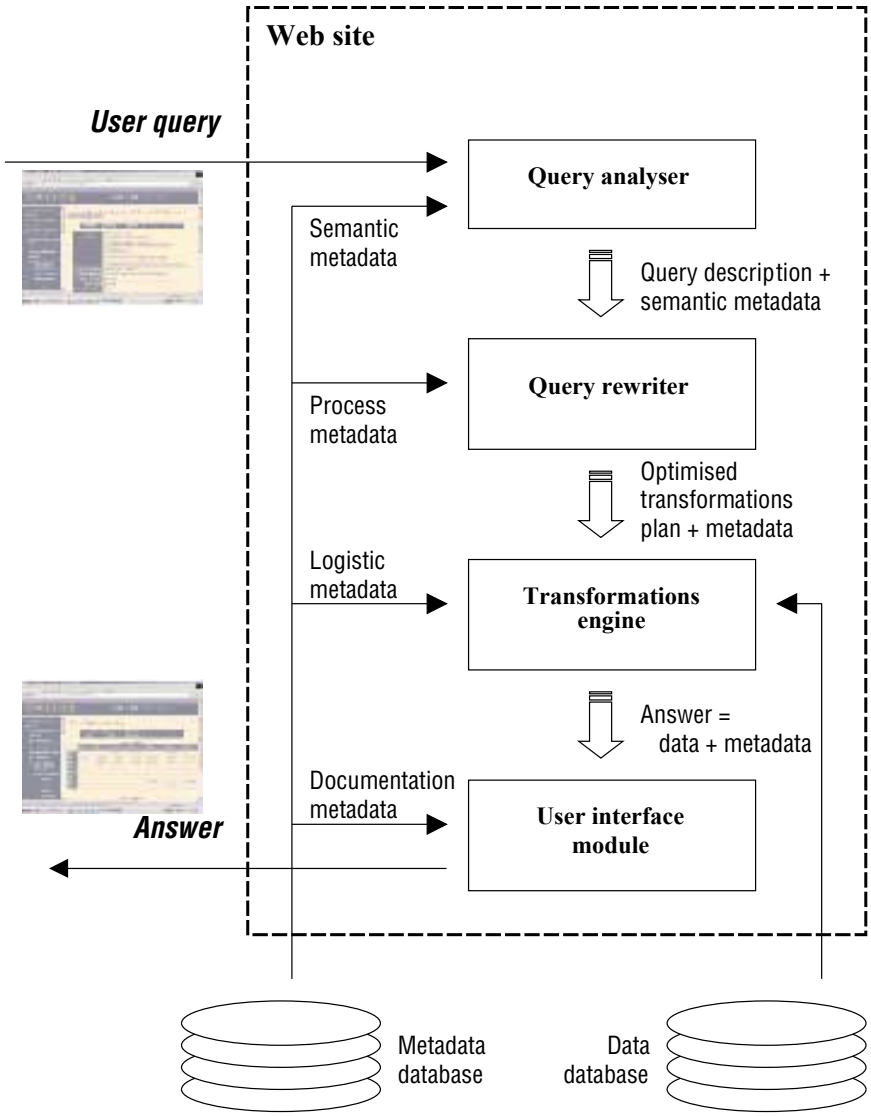


Figure 3. The architecture of a transformations-enabled website.

Figure 3 shows the components of a transformations-enabled site. The user looks at the metainformation that is available at the site (e.g. definitions, nomenclatures, descriptions of surveys, etc.) and finally decides to retrieve a specific table. This table may or may not correspond to an already existing table. Consequently, instead of selecting the table from a list of available ones, the user is presented with a graphical interface that enables him to describe the contents of the required table. This description is automatically coded in a proprietary language (user query) and submitted to the server for further processing.

The server answers the user query by using four distinct modules. The first one (query analyser) receives the user query and semantically checks it against the metadata database. This procedure assures that the query is semantically complete and unambiguous, retrieves all the relevant metadata from the metadata database and finally, forwards the query (together with the retrieved metadata) to the query rewriter. This module is the heart of the system. It uses a set of IT techniques to find an optimised transformations plan. This plan is the actual set of transformations that must be applied onto the existing tables in order to produce the one that the user asked for. Due to the number of available transformations and their properties, usually there will be more than one way of producing the requested table. Therefore, the query rewriter uses heuristics to decide on the optimised plan that seems to consume the least resources and produces results of best quality (e.g. only few missing values, high precision, best match to population coverage, etc.).

The next module that is invoked is the transformation engine. This module takes as input the optimised transformations plan, some additional logistic metadata and runs the actual transformations against the real data. The main feature of this module is that it follows a components architecture (Hatzopoulos et al., 1998) to achieve the flexibility needed to work with the diverse data sources that may be used in a single office (e.g. relational databases, Ms-Excel® sheets, Ms-Access® tables, etc.).

Finally, the last module is responsible for packaging the results into a graph or a pivot table and returning them to the user as an HTML file. This task requires additional documentation metadata that are retrieved from the metadata database. This metainformation is most of the times labels that are used in the tables' headers and footers or in the graphs' axes.

7. Conclusions and suggestions

The idea of automating the processing of statistical aggregates via the use of metadata transformations is a recent one and thus there is substantial lack of documentation. There are only few prototypes, such as Idaresa and IMIM (Eurostat, 1997a), which make trivial use of transformations to produce ad hoc statistical tables (data + metadata). Furthermore, these prototypes can only be used as a proof of concept since they lack many important features such as stability, robustness and compatibility with existing statistical systems, good performance, scalability and user-friendly interfaces.

The metadata models that these systems are using present two major problems. The first one is that these models are completely proprietary designs, which severely reduces the chances of operating them jointly or even complementary to each other. Additionally, there is no metadata-modelling standard to support a unified way of capturing and storing metainformation and consequently, there is no standard to assert the quality of these models.

The second problem of the models is that they are too complex. Usually, it is difficult to convince personnel working in statistical offices to use a metadata model, since the time needed to learn its basics is significant. A partial solution to the problem is to increase the awareness of personnel in the advantages of capturing structured metadata and build suitable user interfaces to hide the complexity of the models. However, this presents difficulties due to the large sizes of the models.

Apart from the previously mentioned IT problems, metadata transformations theory should be further developed. For example, available information is usually incomplete due to missing/vague metainformation or even breaks in time series (Grossmann and Papageorgiou, 1997). Missing values prevent from running bulk imports of existing data into the transformations-aware systems. That is, most of the metadata will have to be imported manually, which is a rather resource-intensive and time-consuming task. Furthermore, breaks in time series endanger the quality of the results obtained through the use of transformations. This happens because a typical transformation plan may contain a chain of eight or more transformations. If an error occurs in early stages, this error will be propagated and probably enlarged by the rest of the chain. However, the user will not double-check the semantics of the result since he believes that the machine generated the data and metadata correctly. Consequently, there is a strong need to find quality metrics for the results obtained via the use of transformations. These metrics will measure the accuracy and completeness of every information produced by transformations. In addition, the effects of error propagation in environments supporting automated productions of tables should be further examined.

8. Acknowledgment

The authors are grateful to an associate editor and two anonymous referees for their extended remarks in a previous version of this paper. Additionally, the authors thank Professors W. Grossmann and K. Froeschl for many helpful remarks.

This work was partially funded by Eurostat's Idaresa and IPIS projects.

9. References

Date, C. J. (1990), *An introduction to database systems,* Volume I, 5th edition, Addison-Wesley, ISBN 0-201-52878-9.

Eurostat (1993), *Statistical metainformation systems*, Office for Official Publications of the European Community, Luxembourg, ISBN 92-826-0478-0.

Eurostat (1994), *Nomenclature for the analysis and comparison of scientific programmes and budgets*, Office for Official Publications of the European Community, Luxembourg, ISBN 92-826-8480-6.

Eurostat (1995), *Research and development, annual statistics 1995*, Office for Official Publications of the European Community, Luxembourg, ISBN 92-827-5109-0.

Eurostat (1996), *Research and development, annual statistics 1996*, Office for Official Publications of the European Community, Luxembourg, ISBN 92-827-8923-3.

Eurostat (1997a), *Development of statistical information systems (DOSIS)*, Office for Official Publications of the European Community, Luxembourg.

Eurostat (1997b), *Eurostat databases, New Cronos 11/1997, CD-ROM version with CUB.X software*, Office for Official Publications of the European Community, Luxembourg.

Eurostat (1998), 'Design of an integrated statistical metainformation system and creation of a CD-ROM on metadata for use in national statistical offices', SUP-COM 1998/LOT 14.

Froeschl, K. A. (1997), *Metadata management in statistical information processing*, Springer, Wien, ISBN 3-211-82987-3.

Ghosh, S. P. (1986), 'Statistical relational tables for statistical database management', *IEEE Transactions on Software Engineering*, 12, pp. 1106–1116.

Ghosh, S. P. (1988), 'Statistics metadata', *Encyclopedia of Statistical Sciences*, Volume 8, pp. 743–746, Kotz, S., Johnson, N. L. and Read, C. B. (eds), John Wiley and Sons, NewYork.

Grossmann, W. (1999), 'Metadata', *Encyclopedia of Statistical Sciences*, update Volume 3, pp. 811–815, 1999, Kotz, S. (editor-in-chief), John Wiley and Sons, New York.

Grossmann, W. and Papageorgiou, H. (1997), 'Data and metadata representation of highly aggregated economic time-series', in *Proceedings of the 51st Session of the International Statistical Institute*, Contributed Papers, Book 2, pp. 485–486.

Hatzopoulos, M., Karali, I. and Viglas, E. (1998), 'Attacking diversity in NSIs' storage infrastructure: the ADDSIA approach', in pre-proc. of *International Seminar on New Techniques and Technologies in Statistics' 98*, pp. 229–234, Italy.

Karge, R. (1998), 'Integrated metadata-systems within statistical offices', in *Proceedings of the Tenth International Conference on Scientific and Statistical Database Management*, Capri, Italy, pp. 216–219.

Kent, J.-P. and Schuerhoff, M. (1997), 'Some thoughts about a metadata management system', in *Proceedings of the Ninth International Conference on Scientific and Statistical Database Management*, pp. 174–185, Olympia Washington.

Layzell, O. and Loucopoulos, P. (1989), *System analysis and development,* 3rd edition. Chartwell-Bratt, ISBN 0-86238-215-7.

Lamb, J. (1998), 'National statistical offices and administrations, and the web: a survey', *Research in Official Statistics*, Vol 1(1), pp. 121–130.

Lenz, H.-J. and Shoshani, A. (1997), 'Summarisability in OLAP and statistical databases', In *Proceedings of the Ninth International Conference on Scientific and Statistical Database Management*, pp. 132–143, Olympia Washington.

Malvestuto, F. M. (1993), 'A universal-schema approach to statistical databases containing homogeneous summary tables', *ACM Transactions on Database Systems*, 18, pp. 678–708.

OMG (1999), *OMG unified language specification*, Object Management Group (OMG) Inc., available on the Internet (http://www.omg.org).

Ozsoyoglu, G., Matos, V. and Ozoyoglu, Z. M. (1989), 'Query-processing techniques in the summary-table-by-example database query language', *ACM Transactions on Database Systems,* 14, pp. 526–573.

Papageorgiou, H., Vardaki, M. and Pentaris, F. (2000a), 'Recent advances on metadata', *Computational Statistics*, 15(1), pp. 89–97.

Papageorgiou, H., Vardaki, M. and Pentaris, F. (2000b), 'Quality of statistical metadata', *Research in Official Statistics*, 2(1), pp. 45–57.

Stonebraker, M. (ed.) (1994), *Readings in database systems,* second edition, Morgan Kaufmann, ISBN 1-55860-252-6.

Sundgren, B. (1991), 'What metainformation should accompany statistical macrodata?', Statistics Sweden R and D Report 1991:9.

Sundgren, B. (1996), 'Making statistical data more available', *International Statistical Review*, 64, pp. 23–38.

Wouter, J., Bethlehem G. K. and Bethlehem, G. J. (1999), 'Integrated statistical data processing', *Research in Official Statistics*, 2(2), pp. 33–45.

W3C (2000), 'Resource description framework (RDF)', available on the Internet (http://www.w3.org/RDF).

Disclosure control methods in the public release of a microdata file of small businesses

Stuart Pursey,

Business Survey Methods Division,
Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6
E-mail: pursstu@statcan.ca

Key words: disclosure control, public use micro data file, confidentiality

Abstract

This paper describes disclosure control methods developed and implemented to release a public-use microdata file of the financial data of small businesses. The paper provides a summary of the data quality of a public-use microdata file based on a previous analysis. The two related issues, disclosure control methods and their impact on the quality of the PUMF, are investigated. The paper offers conclusions on the development of a PUMF for businesses that protects confidentiality and maintains data quality.

1. Background

There are many issues related to disclosure control methodology and many approaches have been adopted. Binder and Desramaux (1995), Eurostat (1996), United States Department of Commerce (1994), and Willenborg and de Waal (1996) provide extensive overviews. This paper describes the methods used to control disclosure and to measure data quality in a microdata file of the financial data of small businesses. The paper also presents conclusions and ideas on the methodology.

In 1996, the Small Business Policy Branch of Industry Canada, the department of the Government of Canada responsible for fostering the economic development of Canada, approached the Small Business and Special Projects Division of Statistics Canada. This branch wanted to establish a publicly available microdata file of small businesses. The file would enable data users the flexibility to analyse the performance of small businesses on a custom basis, provide equal access of data to users, and help improve knowledge of the financial structure of small businesses. The data to support the proposal would come from the taxation data files of the Canada Customs and Revenue Agency (CCRA), the department of the Government of Canada that administers tax legislation.

The Statistics Act of Statistics Canada states that information must not be disclosed in such a way that it is possible to relate the data from a return to any identifiable individual person, business or organisation. In 1971, the Parliament of Canada passed amendments to legislation to allow Statistics Canada access to tax data for statistical purposes. Statistics Canada uses tax data from CCRA as a major source of data for many of its most important statistical

series. The memorandum of understanding, signed by CCRA and Statistics Canada in 1994, describes the protocol for the release of information between the two departments. The understanding is based on the provisions of the Statistics Act, the Income Tax Act, and the Excise Act. They all provide directions to ensure the confidentiality of data. Within Statistics Canada, the Microdata Release Committee reviews and must authorise the public release of microdata. In March 1997, the Microdata Release Committee of Statistics Canada was able to approve the release of a public-use microdata file (PUMF) of financial data of small businesses. This was a first for Statistics Canada since approval for the release of a business microdata file had never before been sought.

1.1. Business microdata

Franconi (1999) and Cox (1995) discuss types of data and their relation to disclosure control methods. They discuss the difficulties in protecting the confidentiality of business microdata. Continuous data are usually extremely skewed because, normally, a few businesses dominate the economy. These businesses are well known publicly and even perturbation of their data (within limits that maintain data quality) does not adequately protect the confidentiality of their data. Some variables are dominated by zero values because few businesses possess a particular financial characteristic and multivariate distributions often show sparse populations of businesses reporting rare combinations of financial characteristics. Generally, it is difficult to protect the confidentiality of businesses within microdata files and maintain excellent statistical quality. But the Industry Canada proposal was about small businesses and it was thought that this sub-population might be homogeneous enough that the typical problem of protecting business microdata might be manageable. For this reason, Statistics Canada began the development of a PUMF for small businesses.

1.2. The data for the PUMF

The data of the PUMF of small businesses originated from a probability sample of tax records drawn from CCRA files for the tax year 1993. Small businesses are unincorporated (T1) and incorporated (T2) businesses, with gross operating revenue between USD 25 000 and USD 5 000 000 in Canadian dollars. The CCRA file (the population file) contains business identifiers (name and address of each business) and two categorical variables that classify businesses into cells. A cell groups together the tax records of businesses with the same T1/T2 category and industry (the four-digit standard industrial classification/establishment (SICe)). This population file also contains several continuous variables: gross operating revenue (for the T1s) or gross operating revenue, depreciation, total equity, total assets, and closing inventory (for the T2s).

The sample file includes the variables of the population file and certain financial data variables captured from the original taxation data forms submitted to CCRA by the businesses. These are variables such as equity, assets (current, fixed, and total), liabilities (current, long-term, and total), profit/loss, revenue, and expenses (cost of goods sold; wages, salaries and

benefits; occupancy costs; and financial costs). There are 25 for the T2s (38 including totals and sub-totals) and 16 for the T1s (24 including various totals and sub-totals). A complete list is shown in Appendix B.

The sample file is drawn using a two-phase sample design. The first phase is a simple random sample stratified by two-digit SICe, province of Canada, three revenue classes, and T1/T2 category. The four-digit SICe is edited (and updated if required) for each record in the first phase sample. The second phase is stratified by the updated four-digit SICe and a sample is selected by simple random sampling within stratum. For unincorporated (T1) businesses that are partnerships, there is the partnership share. T1 returns that are associated with a partnership report for the whole business, not the individual tax filer's share of the partnership. The partnership share provides a means of adjusting without bias the estimates for the repetition of a business in the sample file. The weight for the i^{th} business is:

$$wf_i = wd_i \, w_{T1 partnership\ i} = w_{1st\text{-}phase\ i} \, w_{2nd\text{-}phase\ i} \, w_{T1 partnership\ i}$$

When the partnership share is extremely low, the final sample weight of a business is less than one.

2. Approach to disclosure control

The intruder attempts to link — by accident or by design — the data of any respondent to the identity of that respondent. Disclosure control methods are intended to prevent an intruder from making the correct link. Each data file is unique by its data, users, purposes, and respondents. Consequently, its particular method of disclosure control (with its parameters and thresholds) is unique. Developing a particular technique for a particular file requires making reasonable assumptions about the capability of intruders, then setting disclosure control goals based on the assumptions, then translating the goals to mathematical rules, implementing the rules, and measuring the data quality of the resulting PUMF.

It is difficult to anticipate and fully understand the capability of a potential intruder. Moore (1996) provides a useful investigation of this issue. One might assume that the intruder has

- access to the sample file (before the application of disclosure control methods);
- an understanding of the types and occurrences of non-sampling errors in the sample file and the edit and imputation rules used to deal with them;
- detailed knowledge of the disclosure control methods used to create the PUMF;
- access to a population file where business identifiers reside;
- knowledge that a link between a PUMF record and population record is the correct or incorrect link.

Moore feels that these assumptions are not realistic. He also notes that Muller, Blien and Wirth (1995) share his opinions. Clearly, it is not realistic to assume that the intruder has access to the unperturbed sample: if an intruder has access there is no use and no need for a PUMF. It is most unlikely that the intruder has knowledge of the non-sampling errors and the edit and imputation rules applied to the sample file. Certainly, Statistics Canada keeps the details, most especially the parameters and thresholds, of the disclosure control methods used in a PUMF confidential. (But it is important to say that disclosure control processes have been applied to the PUMF to discourage intruders and to reassure respondents.) For the PUMF of small businesses, we must assume that the intruder may have access to a population file because some government departments, under the Income Tax Act or Excise Act, have access to the file or a variant of it.

This last assumption, access by an intruder to the population file, is perhaps the most important assumption. It meant that we had to protect against an intruder's capability to match a record in the population file (where identification data exists) with its corresponding record on the sample file (where data exists). This assumption by its nature leads to a definite loss of data quality of the PUMF.

In developing this PUMF, it being a first at Statistics Canada for a business file, we emphasised disclosure control — perhaps more than what was required to protect confidentiality — at the expense of data quality. Thus, future development should emphasise finding the edge that minimises the loss in data quality yet provides the correct amount of disclosure control.

From the view of data users, we wanted the PUMF to resemble the original sample file. Thus, other useful methods of disclosure control such as banding continuous variables into categories or even developing a model to generate data were not used for this PUMF. Instead, we began with the original data and found ways to perturb from these data in ways to be described below.

We considered each cell to be a sub-population into itself and thus we ensured that disclosure control was met for each cell.

Four disclosure control goals were set.

A. Ensure that a small proportion of businesses from the population appears on the PUMF and ensure that an intruder cannot know that a particular business is on the PUMF. The purpose is to cast doubt in an intruder's mind that any particular small business even exists on the PUMF.

B. Ensure that each non-zero data value of each continuous variable of the PUMF is perturbed and that an intruder cannot undo the perturbation. The purpose is to cast further doubt in an intruder's mind. If an identification link is made — correctly or incorrectly — the intruder knows that the data of the record is not the data of the respondent — it has been changed.

C. Ensure that a small proportion of PUMF records can be linked correctly to the population file and ensure that an intruder cannot know that a link is correctly or incorrectly made. This is the heart of disclosure control – making it difficult for an intruder to link correctly the data of a business and the identification of the business.

D. Ensure that unique records are removed. They are removed from the PUMF because nothing that maintains reasonable data quality can be done to hide the uniqueness of these records.

Implementing these goals provides an overall level of protection. Thus, there may be a balance among the stringency of these goals that minimises the loss in data quality. During the development of this PUMF efforts were made through trial and error to find a good balance but this issue was not explored thoroughly.

3. Disclosure control

3.1. Modifications to the sample file

The variables that could be used to identify the name or the geographic location of a small business were removed entirely. Removing the geography variable prevents users from having the ability to analyse data by geography. Yet, this seems to have little impact on data quality as an analysis of variance showed geography to be of almost no use in explaining the variability of the gross operating revenue variable, the type of industry variable being much more powerful.

3.2. Goal A: Rates of sampling

Each cell (cells are the businesses with the same T1/T2 category and industry) was modified, if required, so that there were at least ‘s’ records in each cell of the sample file and at least ‘ts’ records in its corresponding population cell. The value of this modification is that an intruder, with access to the population file and the PUMF, finds that few business from the population exist on the PUMF. Thus, for most businesses in the population, a link to the PUMF is not at all possible.

If a cell (population and sample) did not meet these thresholds, then adjoining cells were collapsed using the hierarchical nature of the SICe. This is re-coding the four-digit SICe to its three-digit SICe, and then to its two-digit SICe and even to its one-digit SICe if required) to obtain the desired minimum counts. The modification is useful for data quality and for the implementation of disclosure control rules. We experimented with a variety of minimum sample and population sizes noting that for this CCRA file collapsing by re-coding quickly brings a high sampling rate to a low sampling rate. There is no obvious optimum threshold:

s from 15 to 125 and t from 5 to 20 all seem valid, useful thresholds and thus we chose a convenient pair.

The weight of a record on the sample file provides a strong indication of the probability (and thus the rate of sampling) under which a business was selected. Again, we wanted to ensure that the information available to an intruder conveys that it is rare for a business on the population file to be on the sample file. A business with a weight of one, for example, points to a part of the population that is sampled as a census. Avoiding this difficulty is relatively simple — identify cells that include records with weights that are too small (below a threshold 'Wt_s'), and sub-sample. This process is done separately and independently within each cell. Those that are selected remain in the sample with a fourth weight associated with them, the sub-sample weight, w_s, while those not selected are removed from the sample. The sub-sample design in a cell is stratified simple random sampling where the records are stratified by the size of their original weight. In this way, the strata representing records with small initial weights are the most heavily sub-sampled. In the CCRA file a high proportion of cells required sub-sampling — about 30 % of the T1 cells and 55 % of the T2 cells.

This part of disclosure control has an impact on data quality — the efficiency of the original sample design is undone partially by the sub-sample design. Heavily sampled strata, most certainly take-all strata, cannot exist in a PUMF. In a business survey, one must consider the purpose-— is a PUMF a part of the objective and if so, what sample design is most appropriate? The approach of deriving a PUMF as an add-on to a business survey may not be the best strategy.

The final PUMF weight for a record i is

$$w_i = ws_i\, wf_i = ws_i\, wd_i\, w_{T1partnership\ i} = ws_i\, w_{1st\text{-}phase\ i}\, w_{2nd\text{-}phase\ i}\, w_{T1partnership\ i}$$

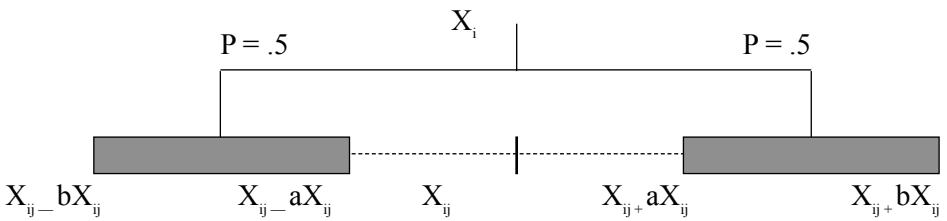
3.3. Goal B: Data perturbation

The implementation of perturbation was achieved in three ways. Random noise was added to each non-zero data value, the three highest data values of each variable were replaced by their average, and all data were rounded to the nearest USD 1 000.

3.3.1. Perturbing data values

We explored a variety of perturbation methods, trying to find one that provided noise yet still had a small impact on data quality. We sorted the data of each variable in descending order and replaced each value by its moving average (spans of 3, 5, and 7 were tried). This perturbation provided minimal loss in data quality but it was not nearly extensive enough to protect the confidentiality of the data. We tried data swapping but found the loss in data quality to be severe.

A traditional approach provides random noise to each datum and so we experimented with a variety of techniques. We provided random noise independently for each datum but it destroyed the mathematical relationships within a record unless the amount of noise was extremely small. We provided a common amount of random noise for each datum within a record (the amount was chosen independently for each record) but this approach moves away too far from the randomness of random noise and thus does not go far enough in protecting confidentiality. Another approach, the one used, is to perturb this way:



This approach provides independent random noise for each datum of a record subject to two constraints:

- within a record the noise is either always positive or always negative;
- the amount of noise is bounded by a minimum of aX_{ij} and a maximum of bX_{ij}.

$Y_{ij} = X_{ij} + \{ B_i(-1,1)\, U_{ij}(a, b)\, X_{ij} \}$ where:
X_{ij} is the datum of the i_{th} record and j_{th} variable
$B_i(-1,1)$ is the random variable, -1 or 1, selected with equal probability
$U_{ij}(a, b)$ is a uniform random variable selected from the interval (a, b) where $b > a > 0$
a is the minimum perturbation as a percentage of X_{ij}
b is the maximum perturbation as a percentage of X_{ij}

Choosing the thresholds (a, b) is based on two issues. In one, the goal is to ensure that each non-zero datum is perturbed so as to confuse the recognition of the business. In the other, perturbation is essential to achieve Goal C. We experimented with a variety of values for (a, b), examining the impact on Goal C. There reached a point where an increase in a and b had little impact on the record linkage results. That is, the proportion of correct links between sample records and population records remained relatively constant. Thus, it was better to abandon increasing random noise and use other techniques to meet the requirements Goal C.

The process has no impact on zeros and it keeps the expected values of a means and total unchanged but increases the variance of the data. Because there is no impact on zeros the pattern of zeros and non-zero remains constant within a record. Thus, rare and unique patterns are visible. This difficulty is addressed in the technique used in Goal D.

Perhaps for a PUMF there is an optimum value for (a, b) based on some objective and subjective analysis but this was not explored deeply. Instead, we selected (a, b) by trial and error, examining values that seemed reasonable to prevent recognition and achieve Goal C.

3.3.2. Averaging the three highest data values of each variable in each cell:

This modification perturbs the three largest data values of each financial variable of each cell, thus dampening the extremes of the data. $Y_{(1)j}$, $Y_{(2)j}$, and $Y_{(3)j}$ are each replaced by $(w_{(1)} Y_{(1)j} + w_{(2)} Y_{(2)j} + w_{(3)} Y_{(3)j}) / (w_{(1)}+w_{(2)}+w_{(3)})$. This has an impact on data quality because it removes some of the natural variability of the data. It does not change the cell's mean and total. Note that the three largest records within a cell vary by variable. This modification is done after the implementation of Goal C.

3.3.3. Rounding all data values to the nearest USD 1 000:

Rounding all data values, Y_{ij}, to the nearest USD 1 000 removes the trailing digits that have little information value. Although this is a perturbation, it is minor in comparison to the other perturbations. But it does provide a visible reminder to intruders and respondents that data modification has been undertaken. This modification is done after the implementation of Goal D.

3.4. Goal C: Linking population and sample records

The implementation of this goal has two parts:

- Calculating proportion of businesses that link correctly between the sample (perturbed data) and population file (unperturbed data).

- Modifying data to reduce the proportion to an acceptable threshold.

The threshold proportion, p, is determined by considering the other thresholds of other goals. As noted earlier there may be a balance among the stringency of all the thresholds that minimises the loss in data quality. During the development of this PUMF we made some efforts by trial and error a good balance but did not explore this issue deeply.

3.4.1. Calculation

Several variables are common to the sample and population file: four-digit SICe, T1/T2 category, gross operating revenue (for the T1s) or gross operating revenue, depreciation, total equity, total assets, and closing inventory (for the T2s). The method an intruder might use to link sample and population records is to attempt an exact match. But the continuous financial variables have been perturbed and so the intruder is able to link correctly cells but not records.

(Since the SICe code may have been updated in the first phase sample or collapsed in achieving Goal A, correct matching cannot always be done. We decided to do what an intruder cannot do: modify the SICe code on the population file as it was modified on the sample file so that all sample and population cells can be matched correctly. This goes beyond what is re-

quired to protect confidentiality for this PUMF because we provided the intruder a power that is not available. Yet, it provides us with an opportunity, within this PUMF, to research this disclosure control methodology more completely since in many sample designs the stratum identifiers are not modified.)

Thus, in our scenario, an intruder is able to link sample and population cells correctly and so we were required to meet Goal C in each and every cell.

Because exact record matches are not possible, another approach of the intruder is to search for close matches, using a powerful method, such as the nearest neighbour approach (for example, this is the approach used in Statistics Canada's generalised edit and imputation system to find donor records in imputation). In nearest neighbour in disclosure control, a sample record is compared to each population record using a distance function. The one with the smallest value of the distance function is the nearest neighbour. This pair, the sample record and its population nearest-neighbour record, might be the correct match of a sample record to itself on the population file.

Playing the role of the intruder, we used a distance function to find the nearest neighbour for each sample record (five variables to match for the T2s but only one variable to match for the T1s). Distance functions include those such as the sum of absolute deviations, sum of squared deviations, and the minimum of the maximum absolute deviation. Transformations to the matching variables (five for the T2 and one for the T1) include those such as the rank value transformation, standardising the data by its mean and standard deviation or by its median and inter-quartile range. An intruder cannot know which distance function and transformation provides the highest rate of correct matches for a PUMF. But as developers of the PUMF, we did what the intruder cannot do: we experimented and used the distance function and transformation with the highest rate of correct matches. This was the rank value transformation. Replace each data value of a variable by its rank divided by the number of data values plus one: $RVT = Rank (X_{ij}) / (N + 1)$. Also, to calculate ranks, combine the sample and population files into one file and then separate them again after ranking. Use, as a measure of nearest neighbour, the sum of absolute deviations. The other distance functions and transformations (even no transformation) provide similar rates of correct matches.

The proportion of correct matches for the T1s was far below the threshold. This happens because there is just one variable to match and random noise itself is sufficient to meet the threshold. But for the T2s, the proportion of correct matches was far above the threshold.

3.4.2. More perturbation to reach the threshold

For each sample record, there is a population record that is the nearest neighbour, the second-nearest neighbour, the third-nearest neighbour and so on. Therefore if the nearest neighbour is the correct link to the population record, it can be made an incorrect link by perturbing the sample record further. The smallest amount of perturbation that changes a correct link into an incorrect link is to replace the data of matching variables of the sample record by the data of the corresponding variables of the second-nearest neighbour population record. With

this perturbation, the nearest neighbour of the sample record becomes a population record that is not the correct link. This type of perturbation is similar to data swapping. It differs because the swap involves two files (population and sample) rather than one file and also the swap is one way (population data remains unchanged).

In a cell with n records and m correct links, if we find m/n greater than p (the threshold for proportion of correct links), then swapping is required for at least (m – p/n) sample records. This is done by selecting (by simple random sampling) m – p/n sample records from the m correctly linked sample records. In doing the data swap, the mathematical relationship of the matching variables to the non-matching variables is lost within the sample record. Thus, the data of the non-matching variables are modified (using prorating) to regain the original relationships. Consequently, all data of the sample record are perturbed, not just the data of the matching variables.

3.5. Goal D: Using a cluster analysis to identify unique records

Some records are so unusual that no amount of perturbation (that maintains data quality) protects them. The most extreme records are identified using principal component analysis with a clustering technique: they appear in the clusters with very few records. The major principal components identify records that are extreme due to large size. The minor principal components identify records that are extreme due to an unusual pattern of data responses. Joliffe (1986) provides a detailed explanation of the use of principal components to detect multivariate outliers. In reviewing the results of this clustering we decided not to remove any records. The most unusual records of this PUMF were the businesses with the largest data values. Yet they were not separated or extreme enough from the rest of the businesses to warrant exclusion from the PUMF.

3.6. Summary of the disclosure control method

Suppose that an intruder declares that a correct link has been made to the identification of a business. Yet, this is unlikely to be true:

- Few businesses from the population are on the PUMF. The sample weight of a PUMF record is high, implying that the business is not unique — instead each business represents many businesses in the population. Further, the intruder cannot know which businesses are on the PUMF.

- If a particular business is on the PUMF, few businesses can be linked correctly, using a powerful linkage tool, to the population file. Further, the intruder cannot know if a link is correct or incorrect.

- If a record is on the PUMF, and if a correct link has been made, then the data values are changed by a minimum (to a maximum) amount. Further data swapping has possibly

added more perturbation, the three largest data values of each variable in each cell have been replaced by their average, any unique records (as determined by a cluster analysis) have been removed, and all data have been rounded to the nearest USD 1 000.

4. Analysis of data quality

4.1. Methodology for an analysis

The movement from sample file to the PUMF changes the original raw data and the statistics created from them. A data-quality analysis may address the original raw data (microdata analysis), the statistics generated from the raw data (macrodata analysis), and the impact at each stage of modification.

Micro-analysis: A measure of quality for a variable is the distance between the 'before datum', X_{ij}, and the 'after datum', Z_{ij}, (the value of X_{ij} after all perturbations are completed). This type of analysis is good for understanding what has happened to the raw data, perhaps without much interest in the purposes of the data. One can create distribution of the measure and estimate its mean, standard deviation, coefficient of variation, median, range, and correlations between variables. The measure can be analysed by SICe, T1/T2 status, variable, and size of business. A typical measure of distance is the proportion that the 'after datum' is from its corresponding 'before datum': $Pd_{ij} = (Z_{ij} - X_{ij}) / X_{ij}$.

Macro-analysis: Data users typically use a microdata file to create descriptive statistics (such as the mean, standard deviation, percentiles and coefficient of variation for a given variable, and the correlations between variables) and often at various domains of the microdata file. A typical measure of the distance between a 'before statistic', S_B, and an 'after statistic', S_A, for a financial variable is the proportion that the 'after statistic' is from its corresponding 'before statistic': $Pd_i = (S_{Aj} — S_{Bj}) / S_{Bj}$.

One can analyse Pd_i by variable, SICe and T1/T2 business status. Both unweighted and weighted analyses can be done. Not using the weights helps in understanding the impact of the disclosure control method on numbers (ignoring the population estimates). Using the weights is best for understanding the impact of the disclosure control method on users' ability to analyse the data of the PUMF and get results close to those obtained from the unmodified sample file.

4.2. Results from a macro quality analysis of Pd_i

In the PUMF, there are 4 208 T1 'sets' of data and 6 775 T2 'sets' of data (where total and sub-total variables are excluded). A set (really the data of a cell disaggregated by variable) contains the data of particular variables (25 for the T2s and 16 for the T1s) of a particular SICe industry (271 for the T2s and 263 for the T1s).

During her cooperative student work term at Statistics Canada, Farr (1997) developed a set of SAS programs to create a database of quality measures for many of the scenarios described in Section 4.1. Metzger (1997), during his cooperative student work term at Statistics Canada, analysed the macrodata quality of the PUMF.

A brief summary of their analyses is shown in Tables 1 to 3 and Tables 1 to 4 of Appendix A. Table 1 shows the percentage of sets where Pd_i is negative, close to zero, or positive. Table 2 provides quality ratings for Pd_i based on the subjective categories shown in the first row. Table 3 provides a statistical summary of Pd_i. Tables 1 to 4 and Box Plots 1 and 2, in Appendix A, are similar but show the quality for two derived variables: total revenue and total expenses.

The T1s statistics have retained much more quality than the T2 statistics and the mean statistic has retained much more quality than the standard deviation statistic. The T1s required much less perturbation than the T2s to meet the disclosure control goals. This is because the intruder would be able to search for a correct link using only one matching variable (T1) rather than five matching variables (T2). Thus, the 'capability' of the intruder is greater for the T2s than for the T1s. The perturbations (especially replacing the three highest value of a variable by its sampling weighted mean) have affected much more negatively the quality of the standard deviation statistic than the mean statistic.

Metzger (1996) was not able to find broad industry groups with consistently high or low quality at their detailed level of SICe code. Yet, there is some consistency in the quality of the variables (except for net operating profit variable, by far the worst quality variable: this is expected because it is extremely sensitive to the smallest imperfections in either of its two components: total revenue and total expenses). Generally, there is a lot of variability in quality, whether examined by variable, by statistic, by industry, or by business status.

Table 1: An analysis of the bias of Pdi over the 16 T1 variables and 25 T2 variables for the weighted mean and the weighted standard deviation

		Percentage negative bias $Pd_i < -0.05$	**Percentage low bias $\lvert Pd_i \rvert \leq 0.05$**	**Percentage positive bias $Pd_i > +0.05$**
Weighted mean	T1	14.19	72.22	13.59
Weighted mean	T2	41.00	26.52	32.47
Weighted standard deviation	T1	39.26	52.00	8.75
Weighted standard deviation	T2	39.01	18.86	42.13

Table 2: Quality ratings over the 16 T1 and 25 T2 variables for the weighted mean and the weighted standard deviation

		Percentage good $\|Pd_i\| \leq 0.05$	Percentage fair $0.05 < \|Pd_i\| \leq 0.15$	Percentage poor $0.15 < \|Pd_i\|$
Weighted mean	T1	72.22	23.65	4.13
Weighted mean	T2	26.52	33.25	40.22
Weighted standard deviation	T1	52.00	30.39	17.61
Weighted standard deviation	T2	18.86	25.90	55.23

Table 3: A statistical summary of Pdi over the 16 T1 and 25 T2 variables for the weighted mean and the weighted standard deviation

	Weighted mean		Weighted standard deviation	
	T1	T2	T1	T2
Number of sets	4 208	6 775	4 208	6 775
Mean of Pd_i	– 0.00575	– 0.00153	– 0.0675	0.0371
Standard deviation of Pd_i	0.106	0.383	– 0.147	0.439
Q1 of Pd_i	– 0.0258	– 0.133	– 0.098	– 0.159
Median of Pd_i	0	– 0.0114	– 0.0245	0.001
Q3 of Pd_i	0.0243	0.0945	0.0103	0.195

5. Conclusions

Controlling disclosure control and maintaining the data quality of a PUMF are the most important issues in preparing a publicly available microdata file for external users.

In this PUMF, disclosure control begins with an assessment of the capabilities of an intruder, followed by the development and implementation of specific techniques of protection. Ensuring that few businesses from the population are on the sample file and the perturbing business data are central to the approach. The sample file is modified so that sampling rates and sampling weights convey to an intruder that the PUMF contains few businesses from the population. The simple recognition of a business is destroyed by data perturbation. More perturbation is provided, if required, to protect against the correct linkage of businesses from the sample and population files. Finally, a cluster analysis is used to identify unique records and, if there are any, they are removed from the PUMF.

Specifying thresholds for the disclosure control techniques are based on a subjective evaluation of the unique characteristics of data, data users, data purposes, and data respondents. Even though each file is unique, some guidance on the values of these thresholds is useful. Generally, we took the approach that each threshold should convey to us (as PUMF developers) and to intruders (even though the values of the thresholds are not made public) that it is discouraging to attempt disclosure. Yet, we did not want thresholds that were so strict that they convey to ourselves or to data users that the PUMF retains little data quality. Recognising that disclosure control takes precedence it is difficult to find a good balance between the two competing needs. We did find it useful to use more than one disclosure control technique — we could trade one off the other to attempt to maintain data quality. As one example, because there is a minimum perturbation, the sampling rate does not have to be 'too low' and conversely because there is a maximum threshold for the sampling rate, the perturbation does not have to be 'too high'.

As one is forced to assume greater capability of an intruder, one loses data quality, sometimes to an unacceptable amount. In this PUMF, we realised that we could not assume that the population file was unavailable to an intruder. This led directly to poor (or at best fair) data quality for the incorporated businesses (the T2s). This was because the intruder had a lot of capability (five matching variables) for the T2s to match the population and sample files. But for the unincorporated businesses there was only one matching variable and we discovered that the original perturbation (to prevent simple recognition of a business) was enough and thus data quality was good. Given that the data quality of the PUMF for the T2s is not good, can the methods of disclosure control be revised, in some optimal way, so that the maximum amount of data quality is retained at no loss in protection? This is an area that would require further research and development.

As discussed at the beginning, it is extremely difficult to derive a public use microdata file of businesses. Yet, an attempt was made because the small business sector does not contain the extreme outliers of the whole business sector. Can the disclosure control approach be taken beyond this sector to the larger businesses? Probably a protected file could be made but most likely (and certainly eventually) it would become a sub-population of the business sector that is not defined and meaningless.

Perhaps there is a sample design that is appropriate for the development of a PUMF for the business sector. Most likely, such a PUMF would contain both individual records and groups of records that are aggregated together. The Small Business Policy Branch of Industry Canada had an explicit definition for the small business sector and this was great advantage for us because this sector overlapped a sector of the business population that is relatively homogeneous.

Under well controlled scenarios — a defined sector of the business sector that is relatively homogeneous and a situation where an intruder has few capabilities (especially no access to a population file where identifiers exists with several variables to match with the sample file) — the development of a PUMF for businesses is a realistic goal. Further research is needed to refine the disclosure control techniques so that data quality is well preserved yet no loss in the protection of respondents' data.

6. References

Binder, D. and Desramaux, L. (1995), 'Confidentiality of statistical data', Statistics Canada internal document.

Cox, L. (1995), 'Protecting confidentiality in business surveys' in Cox, B., Binder, D., Chinnappa, N., Christianson, A., Colledge, M. and Kott, P. (eds) *Business Survey Methods,* pp. 443–473. John Wiley & Sins.

Eurostat (1996), *Manual on disclosure control methods*, produced Helmpecht B. and Schackis, D., Luxembourg: Office for Official Publications of the European Communities.

Farr, H. (1997), 'Examining data quality for a proposed method of creating a small business public use micro data file', cooperative work term report, Statistics Canada and the University of Guelph.

Franconi (1999), 'Level of safety in microdata: comparisons between different definitions of disclosure risks and estimation models', in *Statistical Data Confidentiality*, in Proceeding of the Joint Eurostat/UN-ECE Work session of Statistical Data Confidentiality held in Thessaloniki in March 1999, pp. 25–35, Statistical Office of the European Communities.

Jolliffe, I. T. (1984), *Principal component analysis*, New York: Springer-Verlag.

Metzger, R. (1997), 'Quality analysis of a business microdata file', cooperative work term report, Statistics Canada and the University of Waterloo.

Moore, R. (1996), 'Analysis of the Kim–Winkler algorithm for masking microdata files — How much masking is necessary and sufficient? Conjectures for the development of a controllable algorithm', paper presented at the US Bureau of the Census — Statistics Canada Statistical Interchange, May 13 and 14, 1996.

Muller W., Blien, U. and Wirth, H. (1995), 'Identification risks of microdata', *Sociological Methods and Research*, 24, pp. 131–157.

The Statistics Act (1970, R.S.C. 1985, c. S19) (Canada).

United States Department of Commerce (1994), 'Report on statistical disclosure limitation methodology' prepared by the Subcommittee on Disclosure Limitation Methodology Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Willenborg, L and de Waal, T. (1996), *Statistical disclosure control in practice*, New York Springer-Verlag.

Appendix A

Table 1: Analysis of bias for the weighted mean and weighted standard deviation for total revenue and total expenses

Weighted			Percentage negative bias $Pd_i < -0.05$	Percentage low bias $\lvert Pd_i \rvert \leq 0.05$	Percentage positive bias $Pd_i > +0.05$
Weighted mean	T1	Total revenue	3.42	93.92	2.66
	T1	Total expenses	1.90	97.34	0.76
	T2	Total revenue	42.07	29.52	28.41
	T2	Total expenses	40.59	31.37	28.04
Weighted standard deviation	T1	Total revenue	31.94	58.94	9.13
	T1	Total expenses	39.92	54.75	5.32
	T2	Total revenue	20.30	15.87	63.84
	T2	Total expenses	24.72	18.45	56.83

Table 2: Quality ratings for the weighted mean and weighted standard deviation of total revenue and total expenses

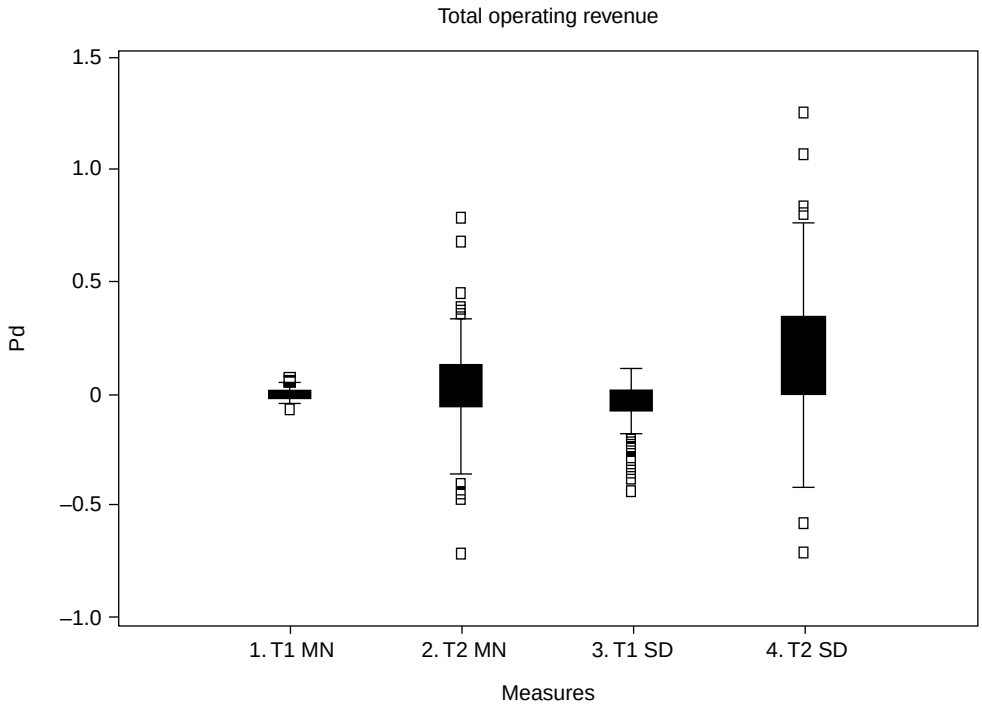
			Percentage good $\lvert Pd_i \rvert \leq 0.05$	Percentage fair $0.05 < \lvert Pd_i \rvert \leq 0.15$	Percentage Poor $0.15 < \lvert Pd_i \rvert$
Weighted mean	T1	Total revenue	93.92	6.08	0
	T1	Total expenses	97.34	2.66	0
	T2	Total revenue	29.52	38.38	32.10
	T2	Total expenses	31.37	39.11	29.52
Weighted standard deviation	T1	Total revenue	58.94	33.84	7.22
	T1	Total expenses	54.75	30.42	14.83
	T2	Total revenue	15.87	28.78	55.35
	T2	Total expenses	18.45	23.62	57.93

Table 3: A statistical summary of Pdi for total revenue

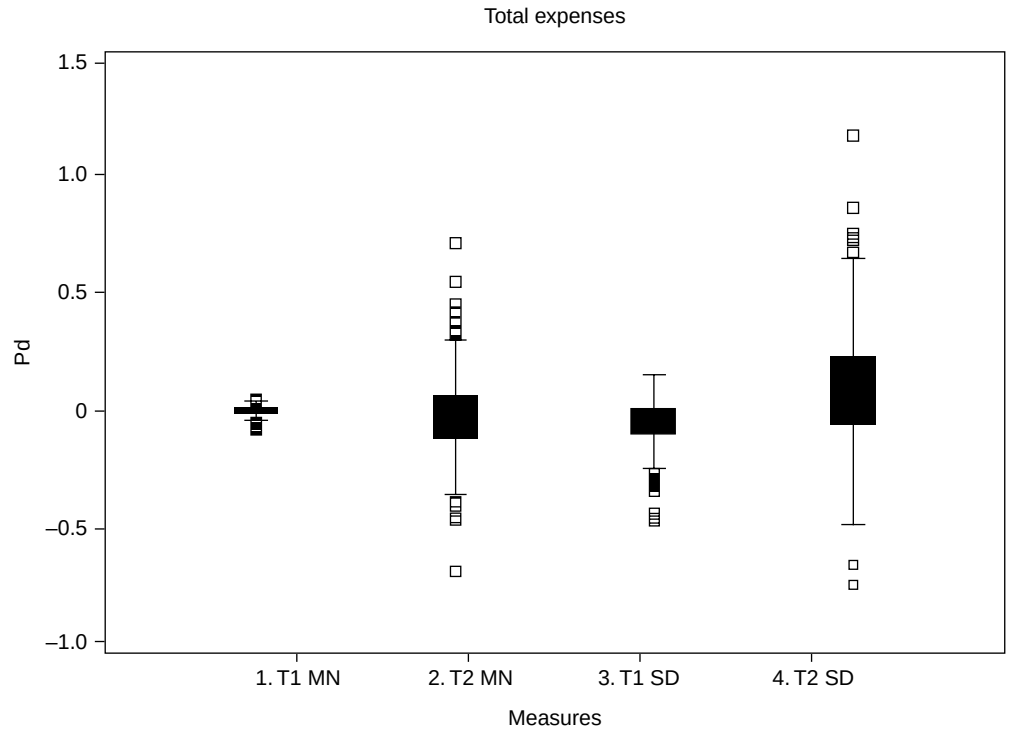
	Weighted mean		Weighted standard deviation	
	T1	T2	T1	T2
Number of sets	263	271	263	271
Mean of Pd_i	– 0.00224	– 0.0165	– 0.0324	0.136
Standard deviation of Pd_i	0.0233	0.175	0.0820	0.264
Q1 of Pd_i	– 0.0154	– 0.103	– 0.0615	– 0.0271
Median of Pd_i	– 0.00147	– 0.0179	– 0.0188	0.119
Q3 of Pd_i	0.0102	0.0766	0.0221	0.302

Table 4: A statistical summary of Pdi for total expenses

	Weighted mean		Weighted standard deviation	
	T1	T2	T1	T2
Number of sets	263	271	263	271
Mean of Pd_i	0.000381	– 0.0196	– 0.0575	0.107
Standard deviation of Pd_i	0.01993	0.170	0.0968	0.272
Q1 of Pd_i	– 0.0103	– 0.113	– 0.0957	– 0.0498
Median of Pd_i	– 0.000665	– 0.0218	– 0.0319	0.0988
Q3 of Pd_i	0.0103	0.064	0.00653	0.237



Box 1: Plots for the weighted mean (MN) and weighted standard deviation (SD) for T1s and T2s



Appendix B: Detail on the variables of the PUMF

There are four types of variables in the PUMF:

Detail variables (25 of them, examples are 'other expenses', 'depreciation', and 'fuel and oil expense').

Subtotal variables, calculated as the sum of detail variables (seven of them: 'cost of goods sold'; 'wages', 'salaries', and 'benefits'; 'occupancy expenses'; 'financial expenses'; 'general expenses'; 'total current assets'; and 'total current liabilities').

Grand total variables, calculated as a sum of subtotal variables and/or detail variables. There are four of them: 'total expenses', 'total liabilities', 'total assets', and 'gross operating revenue'. GOR is included as a grand total variable even though there are no parts to it.

Residual variables: 'profit/loss' (GOR — 'total expenses') and 'total equity' ('total assets' — 'total liabilities').

Balance sheet variables (T2 only)
Total assets
Total current assets
- Accounts receivable
- Closing inventory
- Other current assets

Net fixed assets
Other assets and adjustments
Total liabilities
Total current liabilities
- Current bank loans
- Other current liabilities

Long-term bank loans
Other liabilities and adjustments
Total equity

Categorical variables (T1 and T2)
Industry: SICe at the four-digit level
Business status: T1 or T2

Income variables (T1 and T2)
Gross operating revenue
Cost of goods sold
- Purchases and materials
- Opening inventory
- Closing inventory

Wages, salaries and benefits
- Direct wages, salaries and benefits
- Indirect wages, salaries and benefits

Occupancy expenses
- Depreciation
- Repairs and maintenance
- Fuel and oil
- Utilities
- Rent

Financial expenses
- Interest and bank charges
- Professional fees

General expenses
- Advertising
- Delivery expenses
- Insurance

Other expenses
Total expenses
Net profit/loss

Edisent, automatic filling of electronic questionnaires

Gerrit W. de Bolster, Kees Jan Metz and Jurjen A.T. Piebinga

c/o Statistics Netherlands, Kloosterweg 1, 6412 CN Heerlen, the Netherlands
E-mail: gblr@cbs.nl, kmtz@cbs.nl, jpba@cbs.nl

Key words: administrative burden, electronic (combi-)questionnaire, automatic filling, one-time investment, Edisent, TELER

Abstract

Public administrations in Europe nowadays are requiring more and more information from enterprises. This results in heavy workloads on enterprises for their administrative procedures and the filling of administrative forms. With the Edisent software module, it is possible to reduce the administrative burden. Once the software has been installed at the enterprise, a connection is made with the enterprises administrative information system(s). Electronic questionnaires can be plugged in and after the specification of so-called answer definitions, it is possible to retrieve data items from the administrative information system and use these for automatic filling of the electronic questionnaire. After verification and validation, the data can be sent to the data collector in electronic format. As long as the definitions that are used in the administrative system and the electronic questionnaire do not change, this process of data retrieval, automatic filling and sending can be repeated without further effort.

1. Introduction

In many cases, national statistical institutes (NSIs) send a separate questionnaire to the enterprises for each statistical survey conducted. In order to fill in these questionnaires, these data providers must draw information from several distinct administrative accounts, and then combine and/or recalculate this information in order to bring it into the format that is requested by the data collector. These actions have to be repeated each time the questionnaire is received: yearly, quarterly or even monthly.

Sometimes, the same or similar questions are included in several distinct questionnaires, causing unnecessary overlap. Implementation of modern information technology can help to change this. Instead of starting from the NSI's point of view, from now on, the computerised accounts at the enterprises should be considered as the starting point for data collection. Instead of a questionnaire for each statistical survey, a questionnaire for each computerised account will be developed. The result is an electronic questionnaire that is combining questions for several distinct statistical surveys into one so-called combi-questionnaire.

A precondition for this approach is that the concepts and definitions that are used by the data collector should be attuned to the set of concepts and definitions that the data provider uses. Sometimes, the information required at an NSI can be drawn from the accounts at the enterprise directly, by introducing provisions in the software that the data provider uses, enabling

the data provider to supply the requested information just by pushing another key at the keyboard. In many cases, however, enterprises will use specific, non-standardised concepts and definitions. In order to be able to provide the requested data in an automated fashion in these cases as well, a very flexible tool is needed. For this purpose, the Edisent software module has been developed (Edisent is an acronym for EDI between statistics and enterprises, where EDI means electronic data interchange). This software module must allow for the automated provision of the majority of the data items to be collected, in a format that can be used by the NSI as data collector, thereby reducing the effort to be made by the data provider.

It should be stressed, that it will always be for the data provider to decide whether, and if so, when and under which conditions, the data from his computerised accounts will be made available in this way. This proverbial knife cuts both ways, because for the data collector, this 'EDIfication' will lead to a faster and more consistent way of collecting the data. Thanks to EDI, the effort for processing the data (e.g. validation and editing) can be reduced considerably. It is anticipated that (partly for that reason) the quality of the collected data may be improved, as compared to data supplied via the manual actions that are required for filling in the current questionnaires.

Edisent was developed and tested within the TELER project, an acronym for telematics for enterprise reporting). The project was subsidised by the European Commission under the fourth framework programme. It was carried out by a consortium grouping the national statistical institutes of six (and later eight) EU Member States, representatives of a European federation of accountants and a European professional association for the iron and steel industry.

2. The Edisent module

As far as the NSIs are concerned, a broad definition of electronic data interchange will be used. The main issue here is not the actual sending of the data, but rather the automated collection of the data. The main goal when introducing EDI for the collection of data is to make a substantial contribution to lowering the administrative burden at the enterprise. Using the Edisent module requires a one-time investment for making the connection between the module and the computerised (sub-)accounts at the enterprise. But, once the effort of installation and tuning of the software module has been made, pushing one simple button may do the trick for actually delivering the data to be collected. In cases where concepts and definitions have been standardised, maybe it is possible to buy a standard software module from a software supplier. In that case, the tuning of the Edisent module, as described later, can be omitted.

The new approach that NSIs (and possibly other data collectors) have in mind requires quite a lot of flexibility, because the Edisent module must be linked to the computerised accounts of as many data providers as possible. Since these computerised accounts will differ from one data provider to another, a thorough approach is needed, both in a technical way, a conceptual way and in content.

Technical flexibility is needed to cope with differences in the various software packages used. A conceptual approach is needed in order to reconcile the various sets of concepts and definitions used. As for the content, it is necessary to translate the codes and classifications used at the enterprise into the codes and classifications of the NSIs as data collectors. The Edisent software module takes care of all this. It was initially developed for MS/DOS, at the time of its conception the most used computer platform. During the TELER project, a Windows-version was made available.

3. The technical link

Many different software packages are used today for the computerisation of business accounts. In some cases, standard software packages are used; in other cases, tailor-made software is applied and sometimes a combination of the two. Each software variant has its own way of processing and storing data, but all variants have at least one thing in common: the possibility to produce reports, tables, aggregates, etc., on paper or in a print file of some sort. If it is possible to produce such an ASCII file on an MS/DOS or Windows PC or to transmit it from another computer system to such a PC, then the Edisent module can process this file. A description has to be made for each report, specifying which data items are reproduced in which columns. The Edisent module then can interpret these columns. The only additional requirement is that each 'line' in the report should contain some indication or other through which it can be recognised, for example in financial reports this can be the account number for relevant lines.

4. The conceptual link

All business accounts are built around a consistent set of concepts and definitions. For financial accounts, this may be the general ledger, an enumeration of all entries for booking, with applicable codes and including (sub-)totals. Because in most countries every enterprise may use its self-defined definitions and codes, it is impossible to use a generally applicable classification. Consequently, in such cases, the set of concepts and definitions has to be 'translated' into the concepts of the data collector, as defined in its electronic questionnaire. The Edisent module will offer the possibility to specify a set of simple calculation rules for each question in a questionnaire to define which data item from which ASCII file is to be used and to combine data items from the ASCII file into the data to be collected. The Edisent module will use the column descriptions supplied in the technical link and the formulas for calculating the answers.

5. The link according to content

In addition to the set of concepts and definitions and the accounting schemes (like e.g. the general ledger), various coding lists are also involved, such as article codes, client identifications, supplier identifications and the like. For a data collector, it may be relevant to split

some of the data according to one or more codes, for example a division of sales values according to product codes. Many enterprises and institutes use self-defined codes, even in those countries where a national law defines the accounting scheme as such. Because of this, an option was built into the Edisent module for the translation and/or (re)coding by using tables. The NSI codes can be searched in hierarchically organised lists that are tailored to the branch of the enterprise (classified via standardised NACE codes). The tables for recoding also allow for establishing less obvious links, for example client and supplier identifications can be translated into country codes for supplying data on import or export.

By way of illustration, a simplified fictitious example is given in Figure 1 below. The ASCII file here is called 'report'; it contains the balance sheet of the enterprise from which data items will be derived and used for automatic filling of the electronic questionnaire. In this case, account numbers are used for identification of the lines in the ASCII file: '8000' and '8001' represent sales, '8010' represents discounts.

'L7' and 'K6' represent the article codes, as used in the enterprise's computerised accounts. The data collector for its part is using the product codes '3047062' and '4357781' respectively. The article codes are translated into these product codes during the process, based on a code transformation list (a so-called 'concordance table') that has to be defined within Edisent. In order to define the appropriate values for the questionnaire, a calculation is made: 'debit–credit' for each distinct sales value, 'credit–debit' for the discounts. These calculations are based on user defined rules that are specified within Edisent's so-called 'answer definitions'.

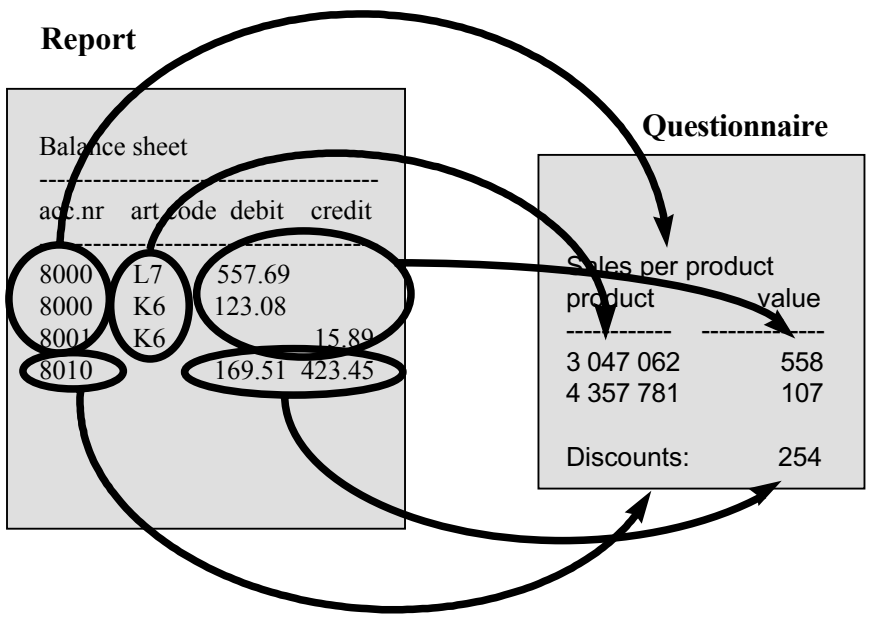


Figure 1: Translation: from report to questionnaire

6. Electronic questionnaires

As mentioned earlier, the information requested by the NSIs will be specified in a specific file called the 'electronic questionnaire'. This electronic questionnaire has to be added to the Edisent software module proper. One Edisent module may contain several electronic questionnaires, for example one for the data required yearly, one for data required quarterly and another one for data required monthly.

If several computerised accounts within the same enterprise are separated physically or organisationally, an Edisent module will be used for each distinct computerised account. The same remark holds if different divisions within the same enterprise use the same computerised account. In the latter case, the electronic questionnaire will be identical. Likewise, administrative offices hired for keeping the computerised accounts may need to use different Edisent modules for the respective branches of activity involved.

In Figure 2 an outline of the Edisent concept is given; the various steps of the process are combined into this one picture, but in practice these steps are executed consecutively. The main functionality of the Edisent module is divided into two parts:

- the initial configuration set-up by which the data types of the information systems at the enterprise are related to the variables (represented by questions asked for in the combi-questionnaire);

- the (periodic) transformation of the data of the enterprise's information systems into the requested messages (answers to the questions).

The Edisent module is suitable for linking to almost any kind of computerised information sub-system. When using the module for supplying data to the NSIs (or other data collectors), only a few steps have to be taken:

1. Apply the software used for the business account(s) to produce reports for the relevant period in ASCII file format, containing the requested data for filling in the questionnaires.

2. Start the Edisent module and fill the electronic combi-questionnaire with these data, after specifying (only once!) the required adjustments.

3. Check the questionnaire after it has been filled in automatically. It can be changed or extended by way of manual data entry (using Edisent as well).

4. After approval, send the data on a diskette or via data communications to the data collector.

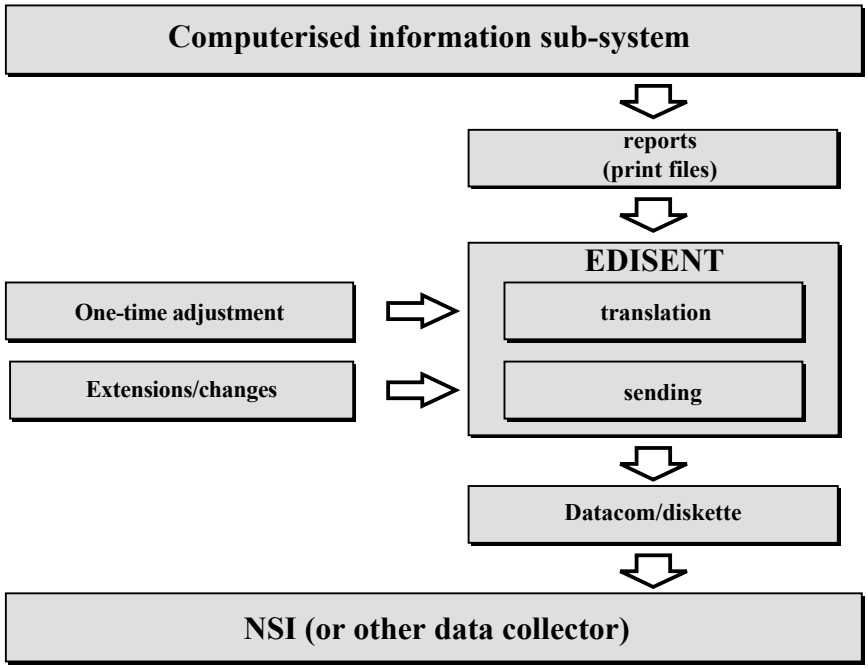


Figure 2: Outline of the Edisent concept

7. Putting Edisent into practice

As mentioned before, there is an Edisent version for both the MS/DOS and the Windows platform. The MS/DOS version is only used by Statistics Netherlands for its own purposes; it currently has been distributed among some 1 200 enterprises in the Netherlands. For the pilots in the TELER project, a 16 bits Windows (3.1 or up) prototype version was made with different, much enhanced functionality. After the end of TELER, Edisent for Windows was adapted: some bugs were solved, some functional improvements were made based on the results of the TELER trials in eight countries and the sending module was rewritten completely and adapted to the Dutch situation. Further improvements on Edisent are being considered.

If feasible, during 2001 the MS/DOS version will be replaced by the Windows version at all enterprises in the Netherlands and subsequently it will be phased out altogether. The Windows version of Edisent has also been used on an ongoing basis by Statistics Finland and SORS (Statistical Office of the Republic of Slovenia). Plans for using Edisent by the other NSIs involved in the TELER project are discussed in the TELER exploitation plan (see 'References').

Recently, it has been decided to use Edisent in the STIPES project (statistical inquiries from popular european software), a project initiated by Eurostat that aims at automating the process of statistical data collection from commercial software packages. In this project, a prototype is developed that derives information from a specific commercial software package's database and use this to fill a selected electronic questionnaire.

8. The future

8.1. Remote metadata control (RMDC)

Electronic questionnaires should be a driving force for achieving better synergy between business information systems and statistical information systems. This is especially true in the case of metadata (classification nomenclatures and other data identification methods). The Edisent concept was thought up in order to bridge the differences in metadata. Using it at the enterprise's side stimulates the inclusion of more statistical metadata within its own computerised systems, thereby making the use of such metadata easier (as a stimulus, not as a necessity).

Remote meta data control (RMDC) describes the way a data collector can manage from a distance the metadata that is stored in an EDI-tool installed at a responding unit. Neither the type of tool nor its manufacturer should be important. For this reason, only open standards are to be selected to use within the RMDC concept.

In the RMDC concept, metadata available on one end of a communication line is combined and packed into a structure before sending it to the other end of the line. On receiving the information, an update of the metadata takes place automatically based on elements in this information. The physical grouping of information elements may differ between sending side and receiving side.

Figure 3 shows this principle.

This principle can be used both for sending metadata from data collector to responding unit and vice versa.

As mentioned above, it is the objective to create a possibility for a data collector to update its own metadata (or business rules) controlling an EDI tool without any interference of the software publisher who has developed this tool. It includes also the possibility that different data collectors can update their metadata controlling one EDI tool independent of each other. The data collectors must be able to do so without the need of developing different internal processes for different 'brands' of EDI tools. The main issue is **standardisation of metadata templates**, not the metadata itself.

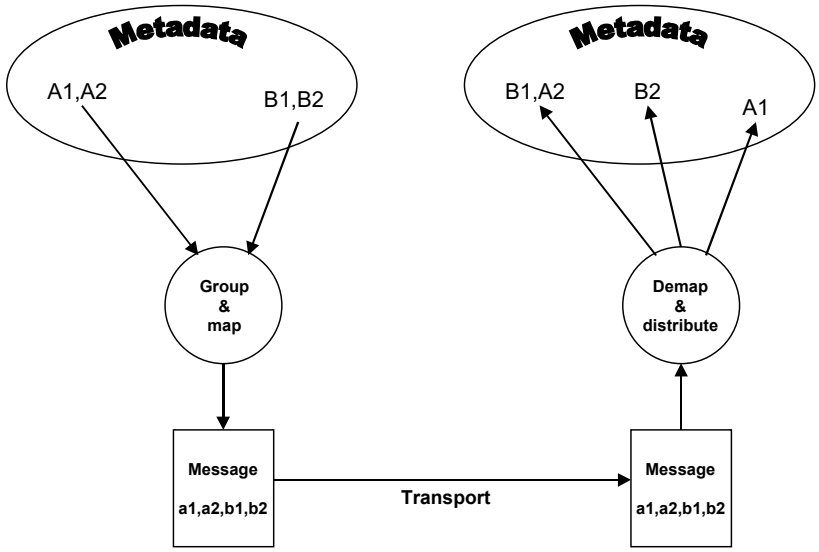


Figure 3: RMDC principle

8.2. EDI sheets

Statistics Netherlands uses several methods to distribute and collect questionnaires:

- questionnaires on paper;
- computer assisted personal interviewing (CAPI) with Blaise;
- computer assisted telephone interviewing (CATI) with Blaise;
- computer assisted self interviewing (CASI) with Blaise;
- electronic questionnaires linked to enterprise information systems (EDISENT, CBS-IRIS).

Many data providers are asking for new ways of interviewing. This caused the starting of a new project named EDI sheets of which the main goal is the development of e-forms for all questionnaires.

The basis assumptions are:

- Each questionnaire will be defined once in a formal language (Blaise).
- A range of e-forms will be generated. This range should suit the needs and the IT-infrastructure of the respondents. The decision of which form to use is up to the respondent.

- After filling in the questionnaires, the results of all appearances are transformed into a standard Blaise file format.
- Only one data-processing system will process the returned data.

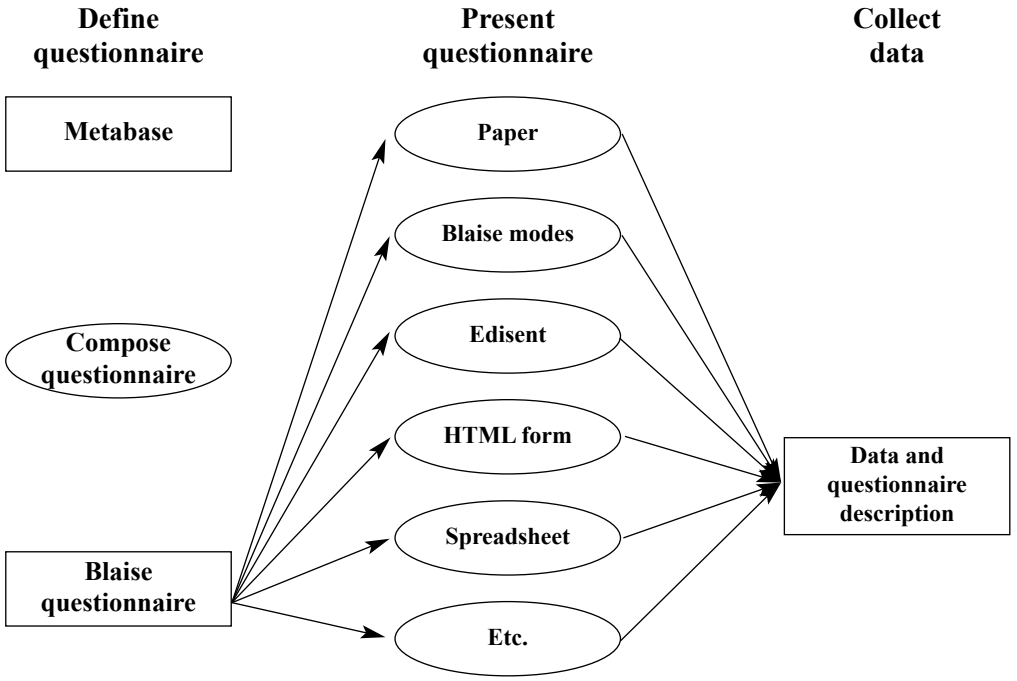


Figure 4: EDI sheets: global process structure

9. Conclusion

The Edisent module is suitable for linking to almost any kind of computerised account. When using the module for supplying data to the NSIs (or other data collectors), only a few steps have to be taken:

1. Apply the software that is used for the computerised account(s) to produce reports for the relevant period, containing the requested data for filling in the questionnaires.

2. Start the Edisent module and fill in the electronic questionnaire with these data, using the one-time tuning.

3. Check the questionnaire after it has been filled in automatically. The results can be changed or extended by means of manual data entry.

4. After approval, send the data on a diskette or via data communications.

10. References

- *TELER on the Internet*
  Edisent is described in much more detail in the TELER documentation. The TELER 'deliverables' are available on the Internet at the following address: http://europa.eu.int/en/comm/eurostat/research/dosis/teler/index.htm
  For first reading, 'Deliverable D08: Final report' can be recommended. Details on possible future use can be found in 'Deliverable D07: Exploitation plan'.

- *Additional info on the Internet*
  Additional information is provided on the Internet in the library section of the CoRD Interest Group of the CIRCA system at the following address: http://forum.europa.eu.int/Public/irc/dsis/Home/main
  In the library section of CoRD, the Edisent user manual and the final report of TELER are available for downloading under 'TELER'. Under 'Inventory of RDC tools and services', up-to-date information on Edisent and contact details are given for those who are interested in a demo version of the Edisent for Windows prototype (that can be used as is, but without any support).

FORUM

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

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

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

Geographic information systems: a challenge for statistical agencies

Mike Coombes

Centre for Urban & Regional Development Studies (CURDS), Newcastle University, Claremont Bridge, Newcastle upon Tyne, NE1 7RU Britain
E-mail: Mike.Coombes@newcastle.ac.uk

Key words: GIS (geographic information system), official statistics, equal opportunities, employment zones, NUTS (nomenclature of territorial units for statistics)

Abstract

The diffusion of geographic information systems is making the handling and analysis of spatial data much easier. One consequence is an increasing demand for spatial data, with users also more likely to expect datasets to be available for a wide range of different sets of areas. Flexibility in data provision runs counter to some traditional practices in the publication of official statistics. This paper argues that it is time to move away from the common practice of rigidly adhering to administrative areas when disseminating datasets. A new role then arises for statisticians: identifying appropriate sets of areas for data dissemination.

1. Introduction

This paper examines certain implications of the diffusion of geographic information systems (GIS) on the publication of spatial statistics, and of official statistics in particular. This issue centres on the selection of the areas for which statistics are presented and analysed. The basic thesis here is in three parts:

- for official statistics, the 'traditional' response to this issue has been to simply present data for a sub-set of the hierarchy of administrative areas;
- recent rapid GIS developments have raised the possibility of a vast increase in the range of areas for which data could be presented;
- there are many contexts in which certain types of area are much more appropriate than others as units for statistical monitoring, and in these cases there remains a task of establishing best practice by identifying the areas which will be the most appropriate.

This paper begins by elaborating on the first two parts of this thesis. The next section of the paper illustrates the need for careful selection of appropriate areas, taking as an example the data required for equal ethnic employment opportunity monitoring. The paper then examines the way in which Britain's 'travel-to-work areas' are defined, because these are one set of ar-

eas specifically designed to be appropriate for certain statistical reporting purposes. The final section of the paper reviews the implications of the thesis as presented.

2. GIS and the availability of spatial data

For any given region of an appreciable size, it would be difficult to list all the different areas for which data users could require such statistics as those from a population census. This unmanageable diversity of demand creates problems for data suppliers, especially when they face financial constraints on their efforts to meet user needs and, in addition, statistical constraints arising from concerns over confidentiality or from the sample sizes of the datasets collected. These statistical constraints usually lead to a clear and defensible population size limit below which a dataset would become unreliable or would put confidentiality at risk. Although numerous different sets of areas will be large enough to meet these constraints, the financial constraints on data providers often mean the dataset is made available for just one or two sets of areas, with the most usual choice being the lowest level in the administrative hierarchy within which all the areas are larger than the minimum size required by that dataset's statistical constraints. Thus, the hierarchy of administrative areas has tended to provide the most commonly used sets of sub-national units for the dissemination of datasets.

This long-standing dependence on hierarchies of administrative areas for data dissemination has two key consequences for users of spatial statistics. The first consequence impacts most strongly on the high proportion of spatial data users who want to analyse 'real world' distributions and patterns, such as those related to the processes of urbanisation: few, if any, sets of areas in the administrative hierarchy provide fully comparable units for spatial analysis. One familiar example is provided by the *Länder* in Germany which range from regions as large as medium-sized countries (e.g. Bayern) to areas which do not even include the whole metropolitan area of a single medium-sized city (e.g. Hamburg). Comparing cities meaningfully requires data on comparably defined meaningful areas, as recognised by the call for definitions of conurbations for the Urban Audit (Taylor et al., 2000). It is inevitable that this problem of non-comparability tends to become still more acute when attempting cross-national comparisons, because even if each country had produced sets of administrative areas which — unlike the *Länder* within Germany — were internally consistent in their definitions, differences between these countries' sets of areas would undermine cross-national comparisons. This is clearly an intractable problem for the cross-national data collation by Eurostat who, having created the NUTS schema as an attempt to align the many different hierarchies of administrative areas in European Union countries (Cardoso, 1997), moved on to assess the potential of GIS to liberate spatial analysis from the limitations of the administrative hierarchy of areas.

It is timely here to reflect on the concerns which lie behind the issue of comparability. For the many patterns and processes which are shaped by the distribution of urban and rural areas, then, a set of areas will only provide meaningful and comparable units for analysis if they are defined in relation to these fundamental features. The need to side-step administrative areas in order to achieve comparability in terms of reflecting local urban geography led

to standard definitions of metropolitan areas in the United States many decades ago, and it has also stimulated numerous efforts to overcome the extra problems facing cross-national comparability in Europe (eg. Pumain et al., 1991). This concern with the geographical meaningfulness of areas will be re-emphasised in following sections of this paper.

The second consequence, of data availability being limited to sets of areas within the hierarchy of administrative areas, is simply that users obtain less information than statistical constraints could permit. The high level of unmet demand for spatial statistics was recently recognised in Britain by the head of the Government Statistical Service who acknowledged that 'the geographic dimensions of UK statistics ... could benefit most from more attention' (McLennan, 1995; p.480). For example, the sample size of the labour force survey leads to statistical constraints in terms of the smallest population areas for which the statistics should be made available. Yet, there are a vast number of possible non-standard sets of areas which may be meaningful to users and which would have large enough populations to meet the dataset's area size requirement. In particular, data for areas spanning national borders will often be of great interest, but any such area will inevitably fall outside the hierarchy of administrative areas enshrined in NUTS because fragmentation along national borders is enshrined throughout the whole hierarchy.

The advent of GIS software has sharpened the interest of data users in this issue. As spatial data processing has been made more widely available, it has led to an increase in the potential demand for spatial statistics; there is also less widespread acceptance that data providers will only release a highly selective sub-set of the information which might have been made available. In short, the diffusion of GIS is undermining users' long-standing (if reluctant) acceptance of data providers' traditional approach of only making available datasets for one or two sets of areas within the hierarchy of administrative areas. Users know that GIS techniques could allow data providers to become much more flexible in terms of the areas for which statistics are provided.

One reason why users are aware of this possible new flexibility is that there are now examples of 'best practice' in data release which do exploit GIS software. A very early example in Britain was the Nomis database (Blakemore and Townsend, 1991) which holds data for areas which are far too small to meet the statistical constraints on data release, but which then allows users to create entirely non-standard areas for which data will then be released if the areas pass the size threshold. New opportunities opened up by the Internet has stimulated numerous innovative data access facilities and related developments by statistical agencies as well as other organisations, particularly in the United States (Rase, 2000). One future role for statistical agencies may then be to create 'data warehouses' which, with GIS-based software, allow users maximum flexibility to aggregate data to non-standard areas in much the same way as Eurostat's GISCO allows the European Union to work on cross-national regions such as the Atlantic Arc (Rase, op. cit.). The datasets available in this environment are best held at the finest possible scale, so that they can be aggregated as accurately as possible to users' areas of interest. There is no need to seek a common 'basic' area for all datasets in any such information system: the overall strategy is compatible with data inputs using administrative areas, postal geography, a grid framework or even microdata (for example, from population registers). The key point here is that an emphasis on data users' needs shifts attention away from constraints imposed by the input data, and towards the potential for removing constraints on the forms of data output.

Thus, the thesis of this paper is that the diffusion of GIS use has transformed the demand for spatial data and, among the 'best practice' cases, is also increasing the flexibility of data release practices. This change could be seen as undermining the statistician's role of deciding upon the appropriate areas for which data should be made available, a decision which in the past usually came down to selecting one tier of the administrative area hierarchy. One concern could be that there is an increased risk of the misuse of statistics: data users could exploit their increased flexibility in the selection of areas for analysis by choosing whichever set of areas most nearly produces their desired results. It is this paper's thesis that the new GIS-facilitated flexibility actually increases the importance of the statisticians' role in the decision over the areas used for data analysis, but that this role is now not to restrict data availability but instead to identify 'standards' in terms of the types of area which are most appropriate for certain types of data analysis. The following sections of this paper provide illustrations of the contexts in which such standards for area definition are required.

3. Equal ethnic employment opportunity monitoring

In the United States, there are well-established procedures for tackling discriminatory employment practices by monitoring and 'benchmarking' with statistics. One of the reasons the 1991 Census White Paper gave for including an ethnic group question in the British census for the first time was that it 'would provide data which would act as a benchmark against which employers and others could measure the success of equal opportunities policies' (HM Government, 1988). How do such policies work in practice? Perhaps surprisingly, the International Labour Office does not provide any guidelines, as it does on gender equality for example (ILO, 1994). In this section of the paper, an existing set of monitoring procedures is examined to see what are the likely requirements for a satisfactory approach to monitoring equal employment opportunity (EEO).

Since 1989, the Fair Employment Commission has had responsibility for identifying and redressing discrimination on the basis of religious affiliation in Northern Ireland (Fair Employment Commission, undated). The Commission obtains monitoring data from employers on the religious affiliation of their employees. However, this information needs to be matched by benchmark data indicating the proportions of catholics and protestants in each local workforce. It is the Commission that determines the boundaries of the relevant local labour market area (LLMA) for each employer. In the Northern Ireland case, the drawing of the LLMA boundaries will be particularly crucial to the results of the labour force profiles. The civil disturbances of recent decades have polarised the two populations so that many people live in areas exclusively housing one community or the other. As a result, the precise LLMA which is drawn round an employer's location will radically affect the religious composition of that area's labour force. Thus, the decision as to which are the appropriate boundaries is critically important to the benchmark statistics set for the employer's recruitment practices.

Similar principles can be applied to policies against ethnic and racial discrimination in the field of employment. In 'monitoring an equal opportunity policy' the Commission for Racial Equality (1988) stated that by comparing their workforce statistics with the relevant external labour market data, employers will be able to establish whether any ethnic and racial minority work-

ers are significantly under-represented or over-represented in any area of their workforce. The CRE guide provides detailed advice for employers on assembling the relevant statistics for their internal workforce, but the experience in Northern Ireland is that the whole outcome will hinge most acutely on obtaining appropriate external labour force benchmark data.

In statistical terms, the idea underlying the measurement of under- or over-representation can be expressed as the aim for the difference between 'observed' and 'expected' values to be minimal. A statistically significant disparity between these values, in relation to the ethnic profile of an employer's workforce, could be seen as *prima facie* evidence that the employer is not complying with the 1976 Race Relations Act's proscription of discrimination in the field of employment (Home Office, undated). Census data is the source for the 'expected' values which indicate ethnic minority groups' share of the LLMA labour force as a whole. Due to the concentration of ethnic minority groups in the inner areas of many cities, their 'share' of such a city's labour force will often depend on how far from the city centre the boundary of the LLMA is drawn. As in most countries, local authority boundaries in Britain are too arbitrarily and inconsistently defined to be relied upon as adequate LLMA definitions, with journey to work patterns invariably crossing these administrative boundaries.

The issues set out here can be examined in a selected sample area centred on the town of Slough (which is located in east Berkshire not far from Greater London's western boundary). Table 1 shows that nearly 23 % of Slough's working residents assigned themselves to one of the ethnic minority groups on their 1991 census forms. Table 1 also shows how this value alters when the area of analysis is altered. The first row covers the population of Slough's local authority area only, and the next three lines in Table 1 extend the area of analysis step-by-step to include those nearby areas with which Slough has strong commuting links. First the neighbouring towns of Windsor and Maidenhead are added to create a grouping which covers East Berkshire. Next the adjacent suburban area of South Buckinghamshire is added, along with nearby Bracknell new town. At the next step, the strong links with west London — and the jobs centred on Heathrow airport in particular — are recognised by adding adjacent London Boroughs which include the airport and nearby suburbs. The remaining rows in the table provide three alternative wider views of the area around Slough itself. The metropolitan dimension is recognised by including Slough within a broad grouping which includes the whole of London and also numerous surrounding areas. A contrast is then provided with data on the county of Berkshire (the unit at the higher level of the 1991 administrative hierarchy which included Slough). Table 1 is completed by data for Britain as a whole.

As stressed earlier, the crucial issue is identifying the appropriate area for the analysis: that is, the LLMA whose data will best represent the labour force for EEO monitoring. For the total workforce, the official set of LLMAs existing in 1991 were known as travel-to-work areas (Dept of Employment, 1984). The grouping of areas 'E. Berks and S. Bucks' (Table 1, row 3) closely approximates the Slough travel-to-work area (as it was in 1991) and so it is highlighted in Table 1 because it can be seen as the appropriate LLMA definition for the total workforce.

Table 1 presents 1991 census data, showing that the choice of area for EEO monitoring has a dramatic effect on the target which would be set for Slough's employers. The ethnic minority

groups make up approaching a quarter of the labour force living in Slough itself, but the equivalent figure is less than one in 12 for E. Berks and S. Bucks (8 %). The simple reason is that the ethnic diversity within the town of Slough contrasts markedly with the overwhelmingly white populations of the neighbouring areas. Table 1 illustrates further the effect of changing the area of analysis through a comparison of the lower rows. The fact that London's ethnic minority groups are by no means all resident in the inner city is shown by the fact that adding the western London boroughs to the E. Berks and S. Bucks area noticeably increases the ethnic minority share of the labour force. The next row adds not only London's inner city but also a wide range of other suburban and home county areas, with the result that the share falls slightly again.

Table 1 also shows that if the analysis options were restricted to the units in the administrative hierarchy, then widening the analysis beyond Slough itself produces a very different result. Slough in fact houses nearly two out of three of all the ethnic minority residents of Berkshire and so an analysis of the whole county gives a value of under 6 % for ethnic minority groups. Table 1 shows that this value is not much higher than that for Britain as a whole, despite Slough being an area with a strong ethnic minority presence.

Three conclusions can be drawn from this example. The first is that the rigid hierarchies of administrative areas will often not provide the most appropriate statistics for users: this was shown by the county data for Berkshire not providing useful information on Slough and its wider context. The second conclusion is that the GIS-enabled freedom to aggregate data into many alternative area groupings can lead to a bewildering range of alternative answers to any question. One consequence could be that some employers might claim that the LLMA boundary of most relevance to them is whichever one yielded the easiest 'expected' share of ethnic minorities in the workforce for them to match with their internal workforce data. The final conclusion then is that the solution must be to explicitly identify, on an objective basis, the most appropriate set of areas for this form of analysis. It is the LLMA which provides the model for such a set of areas, and this paper now turns to the way in which travel-to-work areas (TTWAs) are defined to provide a set of LLMA boundaries for British official statistics.

3.1. Travel-to-work areas

The official monthly series of unemployment data in Britain has provoked a series of debates on statistical issues (Bartholomew et al, 1995). Comparisons between the unemployment rates of towns or cities will only be meaningful if all the areas' boundaries have been defined as consistently and appropriately as possible. The appropriate type of area for this purpose is a set of local labour market areas, because unemployment arises from a shortfall in labour demand relative to the labour supply in that local area. A consistently defined set of labour market area boundaries allows local unemployment rates to be analysed in a more meaningful way than is possible when comparing statistics for areas which, like administrative areas as stressed earlier, have not been consistently defined. This need to compare 'like with like' is particularly acute for unemployment analyses because the less-skilled members of the workforce, who are by far the most likely to be unemployed, tend to only be able to afford to live in a few neighbourhoods; as a result the extent to which the boundary of a city, for example, embraces or excludes these neighbourhoods has a direct compositional effect upon its unemployment rate.

The need for a set of consistently-defined local labour market areas, for the reporting of unemployment rates, provided the 'design brief' which stimulated the development of the procedure for defining TTWA boundaries (Dept of Employment, 1984). The consistency of the areas' definitions is ensured by applying a series of analytical steps, in the form of a computer program, to the published census information on commuting flows within and between over 10 000 small areas. The criteria for the TTWAs were devised to ensure that the resulting boundaries would be as meaningful as possible a set of comparable local labour market areas. A minor concern is with the size of each TTWA's workforce, with a few potentially distinct labour market areas (eg. Scottish islands with smaller populations, such as Mull and Islay) deemed to be below the necessary critical threshold. The reason for such a threshold is that the data series for very small populations are known to be much more volatile than those for larger ones.

The central features of a well-defined local labour market area are that most commuting flows are self-contained within the boundary, and also that those commuting flows make up a pattern revealing substantial integration between the parts of that area (Goodman, 1970). The method of TTWA definition progressively groups the least self-contained areas, choosing whichever grouping leads to a more integrated pattern of commuting flows. In practice, the procedure for defining TTWAs guarantees a consistency of definition in terms of their minimum self-containment and size levels (approximately 70 % and 3 500 workers, respectively). There are also 'target' values (75 % and 20 000) and every TTWA must also satisfy at least one of these, or meet a threshold value on a measure which represents a 'trade-off' between the self-containment and size factors which requires that smaller areas need to be more self-contained (ONS and Coombes, 1998).

The method of definition is a computerised algorithm, which proceeds through four broad stages:

- identify every highly self-contained ward, and every ward with a high rate of in-commuting, and consider each to be the potential 'focus' of a TTWA;
- link together any foci which have high levels of commuting between them;
- gradually assign all the non-foci wards, starting by allocating the ward with the strongest commuting links to the foci (allocating it to that focus — together with any other non-foci wards already allocated to it — with which the ward is most strongly linked), and finishing with the ward with the weakest links overall to other areas (but still allocating it to the one with which it is most strongly linked);
- identify the 'proto TTWA' from Stage 3 which is furthest from meeting the size and self-containment criteria, then re-allocate its constituent wards individually in the same way as Stage 3 (above), and proceed by dealing similarly with whichever 'proto TTWA' is then the furthest from meeting the TTWA criteria, continuing until all the remaining 'proto TTWAs' can be designated as 'draft TTWAs' because they all meet the set criteria.

The technical challenge can be seen from the fact that the analysis has to process a matrix of commuting flows between over 11 000 wards across the whole of Britain. Since the data is not quite a 10 % sample, less than 2 million journeys to work are spread across this matrix of over 100 million cells.

In technical terms, the TTWA method's effectiveness derives from four factors:

- the software can cope with very different sized 'building block' because the early stages of the algorithm group them into localised clusters;
- the method is a multi-stage procedure which avoids the problems which arise when trying to get 'everywhere right first time' and instead it iterates gradually towards a more optimal solution;
- the groupings of areas are not constrained by contiguity — this allows the procedure to always choose the ward which maximises the grouping's integration, although in fact very few areas are grouped non-contiguously in the end because the groupings reflect the commuting patterns which reveal people's reluctance to commute longer distances than are essential;
- the procedure is not rigidly hierarchical, in that two areas which are grouped together at an early stage may then later be grouped into separate areas as part of the 'self optimisation' feature of the algorithm.

The effectiveness of the TTWA algorithm was first shown in 1984 by the vast majority of wards in Britain not needing to be adjusted following their allocation by the computer analysis. Even so, the final 1984 TTWA boundaries did include as a supplementary stage in the research strategy a consultation process on the draft TTWA boundaries emerging from the computerised analysis (Coombes et al, 1986). The responses sought included evidence on major changes to commuting patterns which could be shown to have taken place during the time between the collection of the census data and the undertaking of the consultation stage. The process by which responses were evaluated prevented any changes to the draft TTWAs which would result in defining TTWAs which failed to meet the size and self-containment criteria, according to the best available information at that time. This process was formalised as a multiple criteria analysis in the recent updating of TTWAs based on 1990s data (ONS and Coombes, 1998).

It is notable that once the TTWA procedure had been developed it attracted considerable interest among national statistical institutes facing the same need for consistent labour market area definitions. After the TTWA software had been adopted by ISTAT for their definition of Italian local labour market areas, Eurostat convened a comparative research programme in which several alternative methods were applied to several countries' commuting datasets (Coombes, 1992). The consensus was that the TTWA method had been shown to be the 'best practice' for defining local labour market areas across Europe and so it was the model on which the Eurostat guidelines for their employment zone definitions were based. More recently, a similar review for the US Census Bureau of North

American methods for defining meaningful sub-regional boundaries concluded 'British geographers have developed a more sophisticated computer algorithm for dividing the country into labour market areas (Coombes, Green and Openshaw, 1986)' (Frey and Speare, 1995).

3.2. Review of implications for data providers

This paper has argued that the diffusion of GIS and related spatial data handling facilities has led data users to demand greater flexibility from providers of statistics in terms of the areas for which datasets are made available. There are now several examples of data providers exploiting advanced software to provide flexible forms of access to data, but there are also still many examples of rigid adherence to the traditional approach of only making data available for the lowest set of administrative areas which are all large enough to avoid statistical problems such as those associated with small sample sizes. To continue restricting data release to selected sets of administrative areas is increasingly likely to be criticised as bordering on the suppression of information, when users might reasonably have expected to also have access to data for other areas which meet the dataset's statistical requirements.

From a statistical point of view, data providers may have qualms about the potential consequences of greater flexibility of spatial data release. The ability of users to assemble data for a wide range of different areas may lead to analyses which are seriously flawed, due to the choice of areas for analysis having been made purely to guarantee a particular result. This concern about certain types of area leading to misleading results would be a more robust argument in favour of the traditional approach — that is, only supplying data for selected administrative areas — if it were not clear that the idiosyncrasies of administrative boundaries means that these areas too routinely yield misleading comparisons between different parts of the country. Rather than resisting the mounting pressure for greater flexibility of data release, statisticians could adopt a new role of providing advice or identifying standards which would guide users away from the misuse of spatial data. In particular, new sets of areas might be defined so as to provide the most appropriate units for the analysis of specific datasets.

The discussion in this paper can be summarised as a series of guidelines for the definition of areas used for reporting statistics. The first guideline is a re-statement of the traditional need to identify statistical constraints which set, for example, a minimum population size for data-reporting areas. The example of TTWAs showed that this form of constraint could become quite sophisticated: in the case of the TTWA definitions, a minimum is set for the self-containment level of commuting as well as for population size, and there is then a restricted trade-off allowed between these two factors.

The second guideline is also quite familiar, because it is a call for the maximum level of detail to be provided by releasing data for a set of areas which are as small as is consistent with the first guideline's size constraint. The departure here from the traditional approach for official statistics is to not restrict the choice of areas to those within the hierarchy of adminis-

trative areas. In addition, flexible forms of data release could allow users to define their own areas, with data-access software ensuring that statistical constraints such as those on area size are respected.

The third guideline stems from the recognition that GIS and related developments have led to an irreversible trend towards users' increased choice of the areas by which they analyse datasets. The argument here has been that statisticians should respond by identifying the sets of areas which are the most appropriate for particular analyses. The case of the TTWAs was one which required the data suppliers' statisticians to define a completely new set of areas which then provide a customised set for the reporting of unemployment rate statistics. The crucial point made here was that the most appropriate set of area definitions would reflect the local geography in whichever way is directly relevant to the data for which they will be used. Since the TTWAs are used for the reporting of unemployment rates, the need was for the areas to be a set of LLMAs because it is within labour market areas that labour supply and demand interact to determine the level of unemployment.

Two remaining recommendations concern area-definition procedures. The simpler of the two recommendations is that the methods of area definition used should be defensible as being amongst the most widely respected of their kind. For example, the identification of urban areas might now be expected to exploit satellite imagery to distinguish the full extent of developed land. The other recommendation is that, where possible, the area-definition process should include an element of user consultation. The more complex methods of area definition which are now available are not simply deterministic: there may be a number of solutions which can be shown to meet the set requirements to much the same degree. Thus users' needs could be incorporated into the decision process without departing from those key constraints which will ensure that the defined areas are comparable and meaningful.

The key role of GIS which has been stressed in this paper is that it is tending to stimulate data users' demands. In some cases, GIS can also help data providers to create flexible methods of access to their statistics. Perhaps ironically, the methods and software which are needed for defining new sets of areas are still too specialised to be found in many standard GIS packages.

In most general terms, the challenge to statistical agencies can be seen as the need to respond to new opportunities and user needs whilst adhering to the ethos which was summarised by the 'Fundamental principles of official statistics' (agreed in 1991/92 by the United Nations' Economic Commission for Europe). In practice, these principles had always involved a tension between trying to satisfy all the demands for data which follow from recognising citizens' entitlement to public information, and aiming to retain the public's trust in official statistics by minimising inappropriate usage of those statistics. This paper has argued that the advent of GIS has heightened this tension, and also undermined the customary approach of restricting data availability to those sets of areas in the administrative hierarchy which were considered most appropriate for the presentation of a particular dataset. The alternative way forward set out here is to focus on identifying the most appropriate sets of areas for any particular purpose, with the expectation that more often than not these will not be administrative areas and may need to be defined specifically for that purpose.

Table 1: Employed residents: ethnic* minorities

	%
Slough	22.8
E. Berkshire	11.9
E. Berkshire and S. Buckinghamshire	**8.1**
E. Berkshire, S. Buckinghamshire and W. London	10.9
Greater London and environs	10.2
Berkshire	5.9
Great Britain	4.2

Source: 1991 Population census (ESRC/JISC purchase).
* All groups other than the white majority.
Emboldened area: approximates to the Slough travel-to-work area in 1991.

4. References

Bartholomew, D., Moore, P., Smith, F. and Allin, P. (1995), 'The measurement of unemployment in the UK', *Journal of the Royal Statistical Society*, A158, pp. 363–417.

Blakemore, M. and Townsend, A. (1991), 'Manpower information systems' in Healey M. J. (ed.), *Economic activity and land use: the changing information base for local and regional studies*, Longman, Harlow.

Cardoso, F. (1997), 'The role of Eurostat, its products and its projects', *Cities and Regions*, 10–11, pp. 9–24.

Commission for Racial Equality (1988), *Monitoring an equal opportunity policy* (revised), CRE, London.

Coombes, M. G. (1992), *Study on employment zones*, Eurostat (E/LOC/20), Luxembourg.

Coombes, M. G., Green, A. and Openshaw, S. (1986), 'An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work areas revision in Britain', *Journal of the Operations Research Society*, 37, pp. 943–953.

Department of Employment (1984), 'Revised travel-to-work areas', *Employment Gazette*, September, Occasional Supplement 3.

Fair Employment Commission (undated), *Section 31 review: a guide for employers*, Fair Employment Commission, Belfast.

Frey, W. H. and Speare, A. (1995), 'Metropolitan areas as functional communities' in Dahmann, D. C. and Fitzsimmons, J. D. (eds), *Metropolitan and nonmetropolitan areas:*

new approaches to geographical definition, US Bureau of the Census (Working Paper 12), Washington DC.

Goodman, J. F. B. (1970), 'The definition and analysis of local labour markets: some empirical problems', *British Journal of Industrial Relations* 8, pp. 179–196.

HM Government (1988), *1991 Census of Population* (White Paper, Cmnd. 430) HMSO, London.

Home Office (undated), *Racial discrimination: a guide to the Race Relations Act 1976*, Home Office, London.

International Labour Office (1994), *The ILO at work*, ILO, London.

McLennan, B. (1995), 'You can count on us — with confidence', *Journal of the Royal Statistical Society* A158, pp. 467–489.

Office for National Statistics, Coombes, M. (1998), *1991-based travel-to-work areas*, ONS, London.

Pumain, D., Saint-Julien, T., Cattan, N. and Rozenblat, D. (1991), *The statistical concept of the town in Europe* (Network for Urban Research in the European Community report 0673002), Eurostat, Luxembourg.

Rase, D. (2000), *Commission initiatives to establish a European Geographical Information Infrastructure (EGII)*, Working Paper 6, Conference of European Statisticians, Neuchatel, April, available on the Internet (http://www.unece.org/stats/documents/2000/04/gis/4.e.pdf).

Taylor, D., Bozeat, N., Parkinson, M. and Belil, M. (2000), *The urban audit: towards the benchmarking of quality of life in 58 European cities*, Volume 1, European Commission, Luxembourg.

Generalised software for sampling errors — GSSE *

Stefano Falorsi [1], Daniela Pagliuca [1] and Germana Scepi [2]

[1] Istat, Servizio Studi Metodologici, Via A. Depretis 74/B, 00184 Roma
[2] Dipartimento di Matematica e Statistica, Università Federico II, Napoli
E-mail: stfalors@istat.it pagliuca@istat.it germana@dms.unina.it

Key words: variance estimation, Taylor series approximation, complex designs

Abstract

This paper aims to present the main characteristics of a generalised software implemented in Istat to estimate parameters and compute sampling errors. This software is an important component of a general system developed to standardise the principal operating step related to the sampling strategy of each survey. In particular, this system is constituted of three principal tools: generalised system for the multivariate allocation (GSMA) that deals the problem of defining the best allocation of units into strata, the generalised system of sampling weighting (GSSW) allowing to calculate sampling weights and, finally, the generalised system of sampling errors computation (GSSE) allowing to deal the problem of parameters estimation and the related sampling errors computation. In this paper, we discuss in detail the methodological and functional characteristics of GSSE. Furthermore, we present an example of GSSE application on ISTAT sampling survey data.

1. Introduction

Since the end of 1980s, ISTAT is becoming more interested in developing generalised software (based on general methodologies and a user-friendly interface) because of the increased necessity of standardising important operating steps — such as sample allocation, unit selection, construction of sampling weights, estimation of parameters and sampling errors computation — concerning large-scale surveys sampling strategy. This interest is, in particular, the consequence of the following reasons:

- every year, ISTAT carries out many sample surveys on households and enterprises based on complex sampling strategies;
- it is highly money- and time-spending both to construct ad hoc procedures for each sample survey as well as to modify procedures in order to introduce changes due to the redesign of existing surveys.

Furthermore, it is important to underline that a general theory of estimation, in presence of auxiliary information and a standard method of allocation for more domains and variables of interest, has been developed during the last years.

* This paper is the result of a joint work between the authors. However, Falorsi was mainly responsible for Section 2.1, Scepi for Sections 2.2 and 5 and Pagliuca for Sections 3 and 4. This research was developed by Istat with the collaboration of Piero Falorsi whom the authors thank for useful comments.

As a consequence, generalised software in SAS environment (Falorsi et al.1998, Falorsi. and Rinaldelli, 1998) has been implemented by a group of ISTAT researchers. It is possible to look at this software as a system that includes, in particular, three principal tools and some procedures — programs.

The first tool (GSMA) allows for the definition of the best allocation of units into strata, the second one (GSSW) allows for the calculation of survey weights and the last one (GSSE) allows for the calculation of the estimates of parameters of interest and the corresponding sampling errors. Furthermore, some procedures for solving the problem of unit selection have been developed, both for stratified sampling design, particularly adopted in business survey, as well as for two-stage sampling, adopted for household surveys.

The aim of this paper is to describe the methodological principles and functional characteristics of GSSE. Sections 2.1 and 2.2 we describe the methodological principles behind the software. As the methodologies adopted for the computation of sampling errors are strictly related to the methods adopted for the construction of sampling weights, we also introduce a formal description of the principal methodological characteristics behind GSSW in Section 2.1. In Section 2.2, the methodological principles proper to the computation and presentation of sampling errors are described. In Section 3, we present the main functional aspects of the software: a user-friendly interface to give users an easy way to access data, select variables and input parameters interactively (the most recent version was developed in 1999) and the possibility to choose the outputs. Finally, in Section 4, we present some results related with an application on Istat household surveys data.

2.1. Methodological principles

GSSW software is based on the general theory of calibration estimators (Deville and Sarndal, 1992) and solves the problem of defining the sampling weights for almost all practical situations of estimation concerning ISTAT sample surveys on households and enterprises. The final weights are computed through three steps:

- The computation of direct weights, obtained as the inverse of the inclusion probabilities of the units.
- The computation of non-response corrected weights, or base weights, obtained multiplying direct weights of respondent units by non-response correction factors, calculated for the same units.
- The calculation of final weights: the base weights are corrected multiplying them by post-stratification correction factors. Making the weighted sample distribution of the auxiliary variables conform to external information on these distributions, it is possible to compensate for non coverage and to improve the accuracy of survey estimates.

Let us indicate with U, $U = \{1,..., k,..., N\}$, a finite population of N elements, with s^*, $s^* = \{1,..., k,..., n^*\}$, a random sample of n^* elements, selected from U with the probability

$p(s^*)$, being $s = \{1,..., k,..., n\}$ the sample of n respondent units and $\pi_k = \sum p(s^*)$ the inclusion probability of the generic unit $k \in U$ in the sample. We also indicate with: y_k, the value of the variable of interest y; $\boldsymbol{x}_k = (x_{1k},..., x_{jk},..., x_{Jk})$ the value assumed by the vector $x = (x_1,..., x_j,..., x_J)$ of J auxiliary variables; $z = (z_1,..., z_p,..., z_P)'$ the vector of P auxiliary variables related to the response probabilities of the units, whose values $\boldsymbol{z}_k = (z_{1k},..., z_{pk},..., z_{Pk})$ are known for each unit of the selected sample . Finally, we indicate with r_k the indicator variable that is equal to 1 if the unit is respondent and 0 otherwise.

We want to estimate the total Y of the variable y, given by:

$$Y = \sum_{k \in U} y_k \tag{1}$$

on the basis of the following information:

(a) the values of J + 1 observations $(y_k, \boldsymbol{x}_k)$ for each unit in the sample s;

(b) the values of the J elements of vector $X = (X_1,..., X_j,..., X_J)'$, representing the known population totals of the J auxiliary variables belonging to $\boldsymbol{x}$, being $X_j = \sum_{k \in U} x_{jk}$;

(c) the values of $P + 1$ observations (r_k, z_k) for each unit in the sample s^*.

A general expression of the estimator of the total Y can be the following:

$$\widetilde{Y}_{PV} = \sum_{k \in s} y_k d_k \delta_{ks} \gamma_{ks} = \sum_{k \in s} y_k w_{ks} \tag{2}$$

where $d_k = \pi_k^{-1}$ (per $k = 1,..., n$) indicates the direct weight associated with k^{th} respondent unit, $w_{ks} = d_k \delta_{ks} \gamma_{ks}$ denotes the final weight, being δ_{ks} and γ_{ks} respectively the non-response correction factor, (carried out in step 2), and the post-stratification correction factor (calculated in step 3), for the k^{th} respondent unit.

On the basis of the calibration estimators, using the available auxiliary information for the survey, we utilise GSSW for obtaining**:**

- non-response correction factors δ_{ks} $(k=1,..., n)$ and base weights expressed as $d'_{ks} = d_k \delta_{ks}$ $(k=1,..., n)$, by means of the information previously indicated with (c);

- post-stratification correction factors γ_{ks} $(k=1,..., n)$ and final weights expressed as $d_{ks} = d'_{ks} \gamma_{ks}$ $(k=1,..., n)$, by using the information indicated with (a) and (b).

We have described the estimation procedure considering only the step of final weights computation; the step of non-response corrected weights computation is similar developed.

The set of final weights $\{w_{ks};\ k = l,..., n\}$ is obtained as the solution to a constrained minimum distance problem. The target function is given by:

$$min\left\{\sum_{k\in s} G_k\left(w_{ks}; d'_{ks}\right)\right\} \tag{3}$$

with the following constraints:

$$\sum_{s\in} w_{ks}\boldsymbol{x}_k = \boldsymbol{X} \tag{4}$$

where $G_k\ (w_{ks};\ d'_{ks})$ indicates a distance function between the base weights d'_{ks} and the final weight w_{ks}.

Our aim is to define a set of final weights $\{w_{ks};\ k = l,..., n\}$ allowing both to respect the system of constraints (4) and to modify, as little as possible on the basis of the selected distance function, the set of base weights $\{d'_{ks};\ k = l,..., n\}$. For the solution of the constrained minimum problem, defined by (3) and (4), the distance function, , and its first derivative, $\{G_k\ (w_{ks};\ d_{ks}\}$, must satisfy some regularity conditions (Deville and Sarndal, 1992) assuring that there exists an inverse function $g_k^{-1}\ (.)$ for which $w_{ks} = g_k^{-1}\ (g_k\ (w_{ks};d_k))$.

The solution of the minimum constrained problem (3), (4) consists in the vector $w = (w_{ls}\ ,...., w_{ks}\ ,...., w_{ns})$ obtained by the method of Lagrange multipliers, with (n + J) equations and (n + J) unknown ($\boldsymbol{w}$, λ), where $(\lambda_1,..., \lambda_j,..., \lambda_J)'$ is the vector of Lagrange multipliers.

Starting from the equations Lagrange's multipliers we define the following expression:

$$w_{ks} = d'_{ks}\ F_k\left(\mathbf{x}'_k\lambda\right) \tag{5}$$

where the function $F_k\left(\boldsymbol{x}'_k\lambda\right) = \frac{1}{d'_{ks}}\, g_k^{-1}\left(\boldsymbol{x}'_k\lambda\right)$ represents the correction factor of the base weight, d'_{ks}, and $(x'_k\lambda)$ represents a linear combination of the auxiliary variable vector $\boldsymbol{x}_k$ with the J unknown values of vector λ. The formula (5) is not operative because the numerical values of vector λ are not known. In order to determine λ, after a little algebra, the following system of J equations, with the J unknown quantities λ, is obtained:

$$\mathrm{X} - \widetilde{\mathrm{X}} = = \sum_{k\in s} \mathbf{x}_k\ d'_{ks}\left(F_k\left(\mathbf{x}'_k\lambda\right) - 1\right) \tag{6}$$

in which $\widetilde{\mathrm{X}} = \ \widetilde{X} = \sum_{k\in s} d'_{ks}\mathbf{x}_k$. *If* $F_k\left(\boldsymbol{x}'_k\lambda\right)$ is a linear function of λ, as in the case of the euclidean distance function, we have an explicit solution; otherwise, a numerical solution

can be obtained in an iterative way through the Newton method. Other numerical solutions for specific distance functions are presented in the study of Singh and Mohl (1996). If the vector $\lambda = \lambda$ represents the solution of system (6), it is possible to introduce it in formula (5) and calculate the set of final weights w_{ks} (k = 1, ... , n).

Generally, the euclidean, the logarithmic and the truncated logarithmic functions are the distance functions $G_k(w_{ks}; d_k)$ used to solve a great number of estimate problems arising in large-scale surveys.

GSSW allows the user to utilise these functions as well as other distance functions — described in Singh and Mohl (1996). The euclidean distance function may result in negative or null weights, since the correction factors may vary in the interval $(-\infty, +\infty)$ and usually these weights cannot be accepted in the majority of applications. On the other hand, this function requires a shorter processing time, because solutions are not found in an iterative way. The logarithmic distance function certainly gives positive weights, but they could assume great values that, in general, are greater than those obtained with the Euclidean distance. The advantage of the third distance function, the truncated logarithmic, is that final weights assume values varying in the predefined interval (w_{ks} L, w_{ks} U) and, for this reason, this function is the most used distance function. The values of the parameters L (lower bound of the correction factor) and U (upper bound of the correction factor) are defined by the user. The value of L must not exceed (Verma, 1995) the L_{max} value, which has to be lower than min $\{(X_j / \tilde{X}_j), (j = l,..., J)\}$; the value of U must not be smaller than the U_{min} value, which must be greater than max $\{(X_j / \tilde{X}_j), (j = l,..., J)\}$. The calibration estimator that arises from logarithmic truncated distance function in GSSW is implemented assigning L_{max} and U_{min} as default values for parameters L and U.

It can be shown that all the calibration estimators arising from the different distance functions utilised in large-scale surveys are asymptotically equivalent to the generalised regression estimator (Deville and Sarndal, 1992); furthermore it easy to verify that the generalised regression estimator is the calibration estimator, obtained by using euclidean distance function. For this estimator, asymptotic unbias and consistency properties are verified and linearised expression of sampling variance is known (Särndal et al., 1992).

For the above reasons, the regression estimator plays a central role in the family of the calibration estimators; moreover, this estimator is very important because each form of the regression estimator arises from a particular definition of the regression model joining the study variable y with the auxiliary variables x'. In particular, by the definition of the parameters $\{q_k; k = l,..., n\}$ in the euclidean distance function:

$$G_k\left(w_{ks}; d'_{ks}\right) = \left\{\left(d'_{ks} - w_{ks}\right)^2 / \left(q_k d'_{ks}\right)\right\} \tag{7}$$

and by the definition of the auxiliary information utilised in the construction of the estimator, it is possible to characterise the regression model in terms of the concepts of model type, model level and model group (Estevao et al., 1995). Different definitions of the model type

allow us to obtain different estimators; for example, assuming that $\boldsymbol{x}_k = l/q_k$, ($\forall$ k $\in$ U), corresponding to the model of average, we obtain the Horvitz–Thompson estimator; if $\boldsymbol{x}_k = x_k = l/q_k$, where $\boldsymbol{x}_k$ is a single variable with positive values, corresponding to model of ratio, the estimator of ratio is obtained; if $\boldsymbol{x}_k = (1, x_k)' = 1 / q_k$ corresponding to a simple regression model with intercept, the regression estimator is obtained.

Alternative definitions of sub-populations for which the $\boldsymbol{X}'$ totals are known correspond to different model groups; while the choice of the model level depends on the level (units or groups of units) for which the auxiliary variables are available. Finally, GSSW allows the user to introduce different definitions of model type, model level and model group.

The generalised software for sampling errors, GSSE, allows us to estimate the sampling errors of the parameter estimates (means, totals, and other functions of means and totals) obtained starting from the different calibration estimators implemented with GSSW. Thus, the approximate formulas of the estimated variances of non-linear estimators are obtained by woodruff linearisation method (1971) based on Taylor series expansion. Taking into account the fundamental asymptotic results given in Deville and Sarndal (1992), the linearised expression of generalised regression estimator used for all calibration estimators is the following:

$$\widetilde{Y}_{REG} \cong \widetilde{Y} + (\mathbf{X} - \widetilde{\mathbf{X}})' \mathbf{B} \tag{8}$$

where:

$$\mathbf{B} = (B_1, ..., B_j, ..., B_J)' = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \tag{9}$$

Formula (8) can be written as:

$$\widetilde{Y}_{REG} \cong \mathbf{X}'\mathbf{B} + \sum_{k \in s} d_k \gamma_{ks} e_k^* \tag{10}$$

where γ_{ks} is the correction factor of weight d_{ks} obtained through the Euclidean distance function and $e_k^* = y_k - \boldsymbol{X'}_k \boldsymbol{B}$.

From (10) it is possible to obtain an approximate expression of variance:

$$Var(\widetilde{Y}_{REG}) = \sum_{k \in U} \sum_{l \in U} [\pi_{kl} - \pi_k \pi_l] (d_{ks} \gamma_{ks} e_k^*) (d_{ls} \gamma_{ls} e_l^*) \tag{11}$$

where π_{ks} is the joint inclusion probability in the sample of units k and l. An asymptotically correct estimate of (11) is defined by:

$$\widetilde{V}ar(\widetilde{Y}_{REG}) = \sum_{k \in s} \sum_{l \in s} \left[\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \right] (d_{ks} \gamma_{ks} e_k)(d_{ls} \gamma_{ls} e_l) \tag{12}$$

where $e_k = y_k - x_k' \tilde{\mathbf{B}}$ and $\tilde{\mathbf{B}}$ is an asymptotically correct estimate of vector **B**.

2.2. Methodological aspects of GSSE: computation and presentation of sampling errors

Formula (12) in Section 2.1 is a general expression of variance, effective for sampling with unequal probabilities without replacement. Referring to the complex sampling designs adopted for the surveys carried out by ISTAT, the use of this formula and the calculation of joint inclusion probabilities,, at the different stages of sampling, may produce particularly difficult and computer-intensive results.

As a consequence, GSSE — as many other existing software for sampling variance estimation — adopts approximated formulas for complex designs (one-stage or more-stages sampling designs) involving selection of units with unequal probabilities without replacement.

The adopted formulas are showed below.

In the case of simple random sampling without replacement and for one-stage stratified sampling design with equal probabilities and without replacement, the formula is:

$$\tilde{V}ar\left(\tilde{Y}_{REG}\right)=\sum_{h=1}^{H}\hat{N}_h^2\,\frac{\left(\hat{N}_h-n_h\right)}{\hat{N}_h n_h\left(n_h-1\right)}\sum_{k=1}^{n_h}\left(e_k-\bar{e}_h\right)^2 \qquad (13)$$

where h ($h=1,\ldots,H$) is the stratum index, n_h represents respectively the number of units and the sample size of stratum h, $\hat{N}_h=\sum_{k=1}^{n_h} d_k\delta_{ks}$, and $\bar{e}_h=\frac{1}{n_h}\sum_{k=1}^{n_h} e_k$.

In the case of one-stage or more-stages sampling designs, involving selection of units with unequal probabilities without replacement, the following approximated formula is used:

$$\tilde{V}ar\left(\tilde{Y}_{REG}\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{c=1}^{n_h}\left(\hat{e}_c-\hat{\bar{e}}_h\right)^2 \qquad (14)$$

where n_h represen ts the number of sampling PSU of stratum h, $\hat{e}_c=\sum_{k=1}^{m_{hc}} e_k\, w_{ks}$ (with c indicating the primary stage units (PSUs) index), m_{hc} represents the number of final sampling units belonging to PSU c and $\hat{\bar{e}}_h=\frac{1}{n_h}\sum_{k=1}^{n_h}\hat{e}_c$.

In the case of multistage sampling, the aggregation of all elementary units selected from the same PSU correspond to the utilisation of ultimate cluster method.

When we have only one elementary unit in the stratum or only one PSU in the stratum, GSSE utilises the collapsed strata technique.

Absolute and relative sampling errors are the main statistics used for evaluating sample variability of survey estimates.

Let us indicate with $_d\hat{Y}$ the estimator of the variable of interest *Y* referred to the domain *d*, where by domains we indicate the sub-populations for which the ***X'*** totals are known, corresponding to different model group (see Section 2.1).

The estimate of the absolute sampling error, $_d\hat{Y}$, is given by:

$$\hat{\sigma}(_d\hat{Y}) = \sqrt{\hat{V}ar\,(_d\hat{Y})} \tag{15}$$

where $\hat{V}ar(_d\hat{Y})$ is the sampling variance estimate of $_d\hat{Y}$.

The relative sampling error estimate of $_d\hat{Y}$ *is:*

$$\hat{\varepsilon}(\hat{Y}_d) = \frac{\sqrt{\hat{V}ar(_d\hat{Y})}}{_d\hat{Y}} \tag{16}$$

The sampling errors allow for the evaluation of the accuracy of estimates; moreover, a confidence interval can be constructed, using the absolute error.

For the generic estimate $_d\hat{Y}$, the interval is represented by:

$$Pr\left\{_d\hat{Y} - k\hat{\sigma}(_d\hat{Y}) \quad , \quad _d\hat{Y} + k\hat{\sigma}(_d\hat{Y})\right\} = P \tag{17}$$

In (17) the k value depends on the value of the probability P; i.e., if P = 0.95 then k = 2.

A relative sampling error $\hat{\varepsilon}(_d\hat{Y})$ corresponds to each estimate $_d\hat{Y}$; thus, to adequately use estimates, a corresponding relative sampling error should be shown for each estimate. This is not possible if we consider time, costs and the large number of tables.

Moreover, errors for non-published estimates, which users could calculate, would not be available.

For these reasons, a synthetic presentation of sampling errors, based on regression models, is adopted and implemented in GSSE, considering a relation between the estimates and the related sampling error.

The approach adopted to construct these regression models depends on the type of estimates — frequencies or totals:

- the interpolation of sampling errors referred to frequencies is based on theoretical models: relative sampling errors of frequency estimates are a decreasing function of the estimate values.

- the interpolation of sampling errors referring to totals is particularly complex because there are no theoretical principles for interpolating the estimate sampling errors for these estimates: an empirical approach was adopted in this paper, assuming absolute error as an increasing function of total.

The model used to estimate absolute frequencies, referred to the generic domain d, is expressed by:

$$log\ \hat{\varepsilon}\ (_d\hat{Y}) = a + b\ log\ (_d\hat{Y}) \tag{18}$$

where the parameters a and b are estimated using the least squares method.

GSSE calculates the values of coefficients a and b and the R^2 of the model in (18) (see Table 6 in Section 4) and this information can be used to calculate relative errors for any absolute and relative frequency estimate. GSSE shows the increasing values of the estimates $_d\hat{Y}^k$ *(k=l,...,K)* corresponding to some typical absolute frequency estimates and automatically calculates the related interpolated errors $\hat{\varepsilon}\ (_d\hat{Y}^k)$ (see Table 7 in Section 4). The relative error for a generic absolute estimate can be calculated using easy procedures, as shown in Section 4.

The following model is used to estimate totals, with reference to the generic domain *d*:

$$\hat{\sigma}\ (_d\hat{Y}) = a + b\ _d\hat{Y} + c\ _d\hat{Y}^2 \tag{19}$$

Parameters *a*, *b* and *c* are estimated using the least squares method, considering a large number of points $(\hat{\sigma}\ (_d\hat{Y})\ _d\hat{Y})$. Model (19) is an empirical model. Using the estimated parameters of this model and dividing both members of the model by the estimate value, $_d\hat{Y}$, the following quadratic equation is obtained:

$$a + [b - \varepsilon\ (_d\hat{Y})]\ _d\hat{Y} + c\ (_d\hat{Y})^2 = 0 \tag{20}$$

3. Some functional characteristics of the software

GSSW, requires two input SAS datasets:

The first dataset contains the known totals of each considered auxiliary variable (see Table 1).

Table 1: Variables in the first dataset

Variable Name	Meaning of Variable
Dominio	**Identification code of reference group for the regression model defined to calculate estimator**
TX1	**Known total X_1 of auxiliary variable x_1**
...	**...**
TXJ	***Known total X_j of auxiliary variable x_j***

The second one includes the values that the auxiliary variables assume with reference to each sample unit (see Table 2).

Table 2: Variables in the second dataset

Variable Name	Meaning of Variable
Codice	**Identification code of the generic unit**
Dominio	**Identification code of reference group for the regression model defined to calculate estimator**
COEFF	**Direct weight d_k**
X1	**Value assumed by the auxiliary variable x_1 for the k^{th} unit**
...	**...**
XJ	**Value assumed by the auxiliary variable x_j for the k^{th} unit**

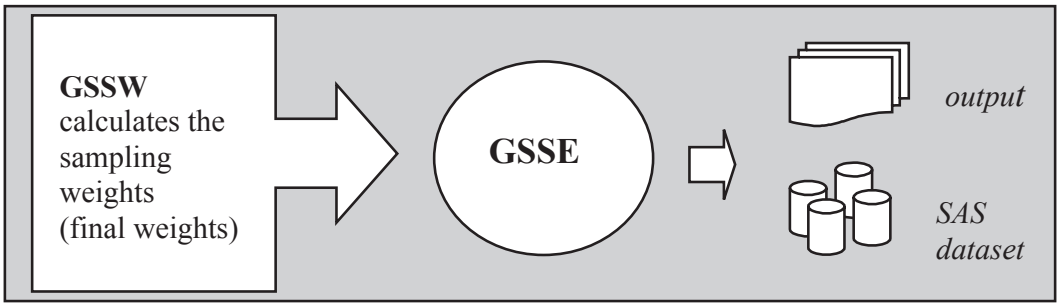
GSSW produces an output containing different information: it provides the base weights, the correction factors and the final weights distributions; furthermore, the differences between estimates calculated with base weights and estimates calculated with final weights are furnished too.

It is important to underline that GSSW allows users to select one of the seven distance function between direct weights and final weights, as already shown in the preceding Section 2.1.

GSSE, calculates the sampling errors on the basis of the final weights obtained by means of GSSW procedure (see Figure1). GSSE solves the problem to calculate:

- domain estimates of population parameters such as totals;
- corresponding estimates of variance (and related statistics such as confidence intervals and percent-relative sampling errors);
- corresponding estimates of statistics giving information on the efficiency of the utilised sampling strategy (sampling design and sampling estimation procedure) such as design effect (deft) and estimator effect;
- regression models for the synthetic presentation of sampling errors.

Figure 1: GSSW, GSSE scheme



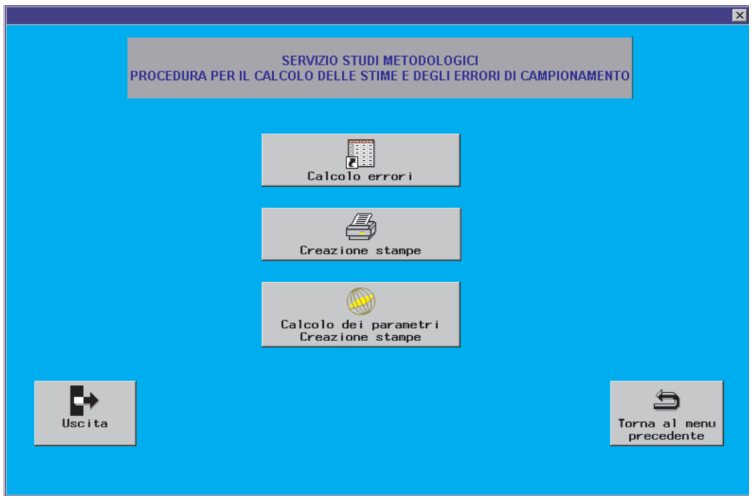
GSSE is developed in SAS environment. The most recent version needs SAS features such:

- SAS language and macro-facility;
- SAS IML language;
- SAS/AF software;
- SCL (screen control language).

GSSE development is at a higher level with respect to the above mentioned software. It presents a user-friendly interface, to give users an easy way to access data, select variables and input parameters interactively. To give an idea of GSSE interface, we present the following figures.

Users can communicate with application through a main window to select input data and open subsequent menus (see Figure 2).

Figure 2: GSSE main menu



First window gives three possibilities:

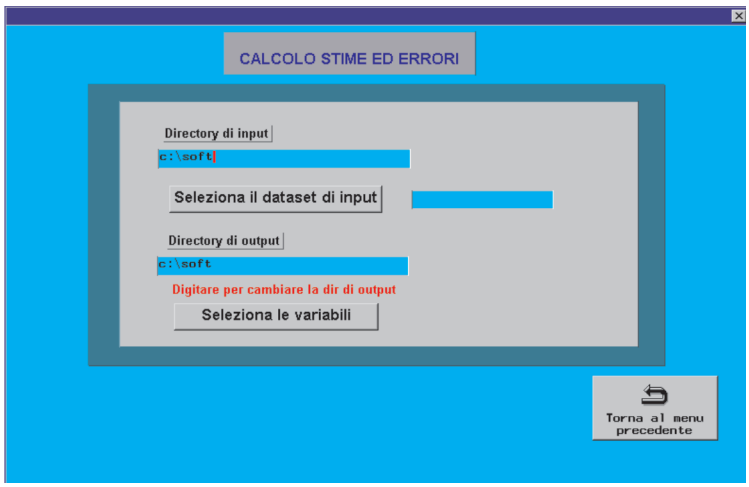
- (1) to evaluate sampling errors and create SAS dataset;
- (2) to print results of a preceding elaboration;
- (1) and (2) simultaneously.

To apply GSSE procedure, the user has to prepare an input SAS dataset, containing the following information for each sampling respondent unit:

- variables of interest (to produce estimates);
- variables defining the utilised sampling design;
- variables defining the type of adopted estimator;
- variables defining estimation domains.

Figure 3 shows the window to use to select data; subsequent windows aid users to select variables.

Figure 3: GSSE dataset window



A window helps users to choose dataset and to define where to place out-put dataset.

Table 3 shows an example of input dataset variables that have to be defined to run GSSE.

Table 3: GSSE input variables

Variable	Type	Meaning
AR	**Design**	**Type of design**
CK	**Estimator**	**Type of estimator**
Codice	**Design**	**Identifier of the final sampling unit**
COEF	**Design**	**D_i = base weight**
Coeffin	**Estimator**	**W_i = final weight**
COM	**Design**	**Identifier of the primary sampling unit**
Dominio	**Estimator**	**Identifier of subgroup for which are known the total of auxiliary variables**
Domstima	**Estimations**	**Planned subgroups for which are requested the estimates**
S1, S2,...	**Estimations**	**Unplanned subgroups identifiers for which the estimates are requested**
Strato	**Design**	**Stratum identifier**
X1, X2,...	**Estimator**	**Auxiliary variables**
Y1,Y2,...	**Estimations**	**Variables of interest**

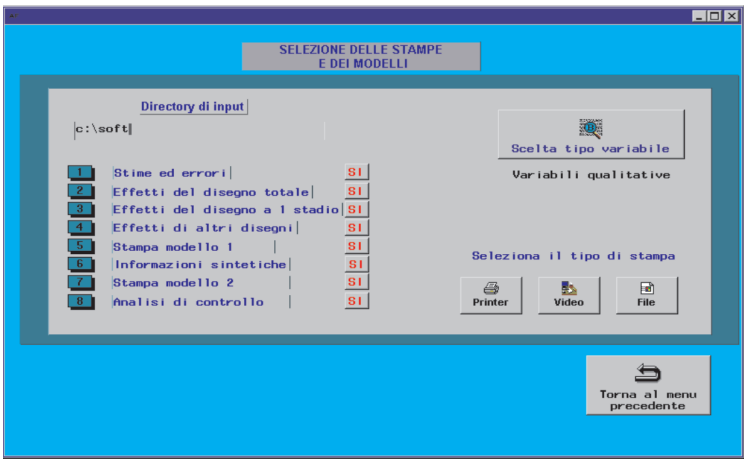
GSSE allows for different specifications of sampling designs, type of estimator and estimation domains and considers explicitly the following sampling designs:

- Simple random sampling with replacement (stratified or not);
- Simple random sampling without replacement (stratified or not);
- Probability proportional to sizes with replacement (stratified or not).

With these designs, the final units may be either cluster or elementary units. With other designs, for example multistage designs, the variance estimates are computed using — by means of an ultimate cluster sampling model — only first stage unit totals without computing variance components for each stage of selection.

GSSE produces a set of outputs. Figure 4 shows the window used to select output.

Figure 4: GSSE output window



User can choose which output wants to obtain (see Table 4).

Table 4: GSSE — Examples of output

Output (a)	Estimates, sampling errors (absolute and relative), confidence intervals both for total population and for planned and unplanned population subgroups
Output (b)	Parameters of the regression models for the synthetic presentation of sampling errors both for total population and for planned population subgroups: $log\ \hat{\varepsilon}\ (_d\hat{Y}) = a + b\ log\ (_d\hat{Y})$ for frequencies, $\hat{\sigma}\ (_d\hat{Y}) = a + b\ _d\hat{Y} + c\ _d\hat{Y}^2$ for totals
Output (c)	Fitted values of sampling errors for pre-assigned values of the estimates both for total population and for planned population subgroups
Output (d)	Statistics for the evaluation of the efficiency sampling strategy (deft and effect of the estimator) both for total population and for planned population subgroups

As shown in Table 4, with GSSE it is possible to obtain estimates and sampling errors and the related confidence intervals; in fact, GSSE software constructs confidence intervals, using the absolute error. Other PC software allows for the calculation of this information. Furthermore, in addition to the computation of the general information about estimates, it is possible to obtain a synthetic presentation of sampling errors, based on regression models, and it is the possible to evaluate the efficiency of the sampling strategy.

To show in detail the use of the software, in Section 4 we present an outcome obtained with a real application on Istat households survey data.

4. Some results of GSSE application to Istat survey data

GSSE software is currently used by Istat to calculate sampling errors concerning the estimates of the principal large-scale sampling surveys on social areas, where resident households and their members are the examined target population ([2]). The 'labour force' survey, the 'consumer expenditure' survey and the 'multipurpose' survey are some examples of these surveys. In particular, the multipurpose survey has various aims, related with different aspects of social life. This survey is carried out yearly during a week in November and the reference period is the year (12 months) before this week.

In the multipurpose survey, estimates are calculated yearly for both households and individuals.

In this section, we present some results concerning the households and, in particular, we show some tables obtained with the use of GSSE.

To give some information about this survey, principal characteristics are presented here; it is important to underline that these characteristics are the general characteristics of Istat surveys on households:

(a) Reference domains

The reference geographical domains for estimates are:

— the whole country;

— five main geographical areas (north-east, north-west, centre, south, islands);

[2] A household is a group of people living together and related or connected by marriage, kinship, affinity, adoption, etc. (Istat, 1994).

— geographical regions;

— six areas defined by the social and demographic characteristics of municipalities.

Moreover survey has the aim to furnish estimates referred to 18 areas that are particularly interesting for statistical analysis: urban areas: A_1–A_{13} major urban areas; A_{14} satellite municipalities of major urban areas; non-urban areas: A_{15}–A_{18}

(b) Sample design

The sample design is a two stage stratified design: for each selected sampling unit, corresponding to a household, all the members (cluster) of the households are surveyed.

In 1998, the survey was based on a sample of less than 29 000 households and 900 municipalities.

(d) Estimates

Estimates are calculated for both households and individuals, using a calibration estimator (see Section 2.1), that is the standard estimator adopted in Istat surveys on enterprises and households. To obtain the final weights, in order to compensate for non-coverage and to improve the accuracy of survey estimates, we use 18 auxiliary variables, related with the regional distribution of population for sex and age in the areas A_1–A_{18}.

(e) Sampling errors

Absolute and relative sampling errors are the main statistics to evaluate sample variability of survey estimates. Tables 6, 7 and 8 (outputs *b*, *c* and *d* presented in Table 4) were directly obtained through the application of GSSE to the multipurpose survey (data of 1998).

Looking at these tables, it is possible to focus our attention on the following point: through GSSE software, in addition to the computation of the general information about estimates, there is the possibility to evaluate the sample variability of survey estimates — Tables 6 and 7 — and the efficiency of the sampling strategy — Table 8 — at the same time; these last two tables are particularly interesting. In fact, Table 6 shows the values of coefficients a and b and the R^2 of the model $\log \hat{\varepsilon}\,(_d\hat{Y}) = a + b \log (\hat{Y}_d)$ for frequencies, used to interpolate the sampling errors of frequencies, for each geographical area A_1–A_{18} and Italy (A_1–A_{18} and Italy represent different domains *d* — see Section 2.2); the information in Table 6 can be used to calculate relative errors for any absolute (and relative) frequency estimate.

Table 6: Values of coefficients a, b and of R2 of functions used to interpolate sampling errors for absolute frequency estimates by geographical areas A1–A18 and the total country (Households — 1998).

Geographical domains	*a*	*b*	R^2
1	7.40174	– 1.18511	97.37
2	7.42747	– 1.20291	97.39
3	8.11784	– 1.18609	97.43
4	6.68539	– 1.20909	96.74
5	7.12274	– 1.21635	97.09
6	7.13380	– 1.19944	97.39
7	7.00660	– 1.21413	97.30
8	8.19545	– 1.17410	97.63
9	6.94257	– 1.16819	96.51
10	6.60084	– 1.20009	97.10
11	7.82769	– 1.18149	95.76
12	6.36232	– 1.21853	97.01
13	6.23612	– 1.21523	97.48
14	7.59659	– 1.05876	91.94
15	6.90615	– 1.01651	91.50
16	7.72493	– 1.06929	93.71
17	7.88976	– 1.07322	93.99
18	8.46157	– 1.14384	95.24
TOTAL	**8.03119**	**– 1.09212**	**95.47**

In figure 5, we show the information that can be found in Table 7: the first column presents some increasing values of the estimates $(\hat{Y}_d)$ (k = 1, … ,K) and the second column shows the related interpolated errors, for each area.

Figure 5: Interpolated values of the percent relative sampling errors, referred to some typical frequency estimates (columns in Table 7)

Estimates (%)	Area_x Relative errors (%)
$\hat{Y}^1$	$\hat{\varepsilon}(\hat{Y}^1)$
…	…
$\hat{Y}^k$	$\hat{\varepsilon}(\hat{Y}^k)$
…	…
$\hat{Y}^K$	$\hat{\varepsilon}(\hat{Y}^K)$

To calculate sampling errors more easily, it is possible to use data in Table 7 (results are less accurate than those obtained using the model $\log \hat{\varepsilon}(\hat{Y}_d) = a + b \log(\hat{Y}_d)$). In Table 7, we present the interpolated values of the percent-relative sampling errors referring to the geographical non-urban areas A_{15}–A_{18}. With this information, the relative error for a generic absolute estimate can be easily obtained using two easy procedures. The first procedure is based on the identification, in the first column, of the estimate closer to the estimate $\hat{Y}_d$; the relative corresponding error $\hat{\varepsilon}(\hat{Y}_d)$ is the value reported on the second column. The second procedure determines the sampling error of the estimate $(\hat{Y})$, through the following expression:

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) + \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^{k})}{\hat{Y}_d^{k} - \hat{Y}_d^{k-1}} (\hat{Y}_d^{k} - \hat{Y}_d^{k-1}) \qquad (21)$$

where: $\hat{Y}_d^{kl}$ and $\hat{Y}_d^{k}$ are the estimate range in which $\hat{Y}_d$ is included and $\hat{\varepsilon}(\hat{Y}_t^{kl})$ and $\hat{\varepsilon}(\hat{Y}_d^{kl})$ are the corresponding relative errors from the table.

Table 7: Interpolated values of percent relative errors of absolute frequency estimates by geographical non-urban areas A15–A18 (A15, municipalities up to 2 000 inhabitants; A16, municipalities 2 001–10 000 inhabitants; A17, municipalities 10 001–50 000 inhabitants; A18, municipalities of 50 000 inhabitants and over) (households, 1998)

	Area A15		Area A16		Area A17		Area A18	
Estimates (%)	Estimates (absolute value)	Relative errors (%)	Estimates (absolute value)	Relative errors (%)	Estimates (absolute value)	Relative errors (%)	Estimates (absolute value)	Relative errors (%)
0.10	1 367.78	80.49	5 164.15	49.24	4 980.47	53.61	3 474.72	64.90
0.50	6 838.91	35.52	25 820.77	20.83	24 902.36	22.60	17 373.62	25.85
1.00	13 677.81	24.97	51 641.54	14.38	49 804.73	15.58	34 747.25	17.39
2.00	27 355.63	17.56	103 283.08	9.93	99 609.46	10.74	69 494.50	11.70
3.00	41 033.44	14.29	154 924.61	7.99	149 414.19	8.64	104 241.75	9.28
4.00	54 711.26	12.34	206 566.15	6.85	199 218.92	7.41	138 989.00	7.87
5.00	68 389.07	11.02	258 207.69	6.08	249 023.65	6.57	173 736.25	6.93
6.00	82 066.88	10.05	309 849.23	5.52	298 828.37	5.96	208 483.50	6.24
7.00	95 744.70	9.29	361 490.77	5.08	348 633.10	5.48	243 230.75	5.72
8.00	109 422.51	8.68	413 132.30	4.73	398 437.83	5.11	277 978.00	5.29
9.00	123 100.33	8.18	464 773.84	4.44	448 242.56	4.79	312 725.24	4.95
10.00	136 778.14	7.75	516 415.38	4.20	498 047.29	4.53	347 472.49	4.66
15.00	205 167.21	6.31	774 623.07	3.38	747 070.94	3.64	521 208.74	3.70
20.00	273 556.28	5.45	1 032 830.76	2.90	996 094.58	3.12	694 944.99	3.14
25.00	341 945.35	4.86	1 291 038.45	2.57	1 245 118.23	2.77	868 681.23	2.76
30.00	410 334.42	4.43	1 549 246.14	2.33	1 494 141.87	2.51	1 042 417.48	2.49
35.00	478 723.49	4.10	1 807 453.83	2.15	1 743 165.52	2.31	1 216 153.73	2.28
40.00	547 112.56	3.83	2 065 661.52	2.00	1 992 189.16	2.15	1 389 889.98	2.11
45.00	615 501.63	3.61	2 323 869.21	1.88	2 241 212.81	2.02	1 563 626.22	1.97
50.00	683 890.70	3.42	2 582 076.90	1.78	2 490 236.46	1.91	1 737 362.47	1.86

Table 8 contains some other information that can be obtained with GSSE; in this table, we present the values of two statistics that are really interesting, as they allow to evaluate the efficiency of the sampling strategy: the 'deft' and the 'effect of the estimator'. Deft values, for example, are obtained automatically from the software: GSSE in fact, for each geographical domain, computes the ratios between the standard errors calculated with the effective design and the standard error calculated with a simple random sample.

Table 8: Statistics for the evaluation of the efficiency sampling strategy (deft and effect of the estimator) by geographical areas A_1–A_{18} and total country (households, 1998)

Geographical domains	Deft	Deft max.	Eff. stim.	Eff. stim. Max.
1	0.88	1.11	0.88	1.06
2	0.88	1.11	0.87	1.05
3	0.89	1.09	0.88	1.03
4	0.85	1.16	0.84	1.03
5	0.88	1.18	0.86	1.04
6	0.87	1.16	0.86	1.08
7	0.87	1.20	0.84	1.07
8	0.90	1.09	0.87	1.03
9	0.98	1.41	0.88	1.14
10	0.91	1.17	0.88	1.05
11	1.61	2.12	0.86	1.12
12	0.49	0.67	0.88	1.02
13	0.90	1.16	0.86	1.05
14	0.99	1.88	0.86	1.15
15	1.11	2.25	0.81	1.19
16	1.19	2.35	0.86	1.13
17	1.16	2.09	0.84	1.07
18	1.00	1.76	0.85	1.17
TOTAL	1.21	2.01	0.86	1.08

5. Conclusion

The main interest of GSSE software consists in standardising operating steps proper to the sampling strategy of each survey. In particular, this software is currently used by Istat in order to calculate absolute and relative sampling errors obtained starting from different calibration estimators.

The computation of sampling errors is obtained according to area and to sampling design. In fact, the computation of sampling errors involves, generally, three factors jointly:

- the definition of sampling design;
- the use of the auxiliary information;
- the choose of the estimator utilised in the process.

GSSE aims to integrate the previous three factors in a unique context allowing users to deal with a wide set of different sampling strategies.

These strategies are implemented by Istat for principal large scale sampling surveys both on social areas as well as on business; the 'Labour force' survey, the 'Consumer expenditure' survey and the 'Small enterprises survey' are some example.

In most cases, Istat sampling surveys on households are based on a composite type design.

Within a given territorial domain of study (e.g. geographical regions), the municipalities are divided into two area types: the self–representing areas (SRAs), consisting of the larger municipalities, and the non self-representing areas (NSRAs), consisting of the smaller municipalities.

In SRA, stratified cluster sampling is adopted. Each municipality is a single stratum and the primary sampling units (PSUs) are the households selected with equal probability by means of a systematic sampling; all members of each sample household are interviewed.

In NSRA the sample is based on a two-stage sample design stratified for PSUs. The PSUs are the municipalities, while the secondary sampling units (SSUs) are the households. The PSUs are divided into strata having approximately equal populations. One or two PSU samples are selected from each stratum without replacement and with probability proportional to size (total number of persons). The SSUs are selected without replacement and with equal probabilities from the selected PSUs independently. All members of each sample households are enumerated.

In contrast, the principal business surveys by Istat are based on a simple random sampling. Within a given territorial domain of study the enterprises are stratified on the number of persons employed and the branch of economic activity of local units.

GSSE shows the useful characteristic of considering simple designs as well as composite type designs. For this reason, the input dataset contains a variable defining the type of design for each sampling unit. Some statistics are produced by GSSE for the evaluation of the efficiency of the chosen sampling strategy.

Furthermore, GSSE allows to define models for synthetic, and not more expensive, presentation of sampling errors based on a particular regression models.

Finally, it is very important to underline the user-friendly interface of GSSE that allows users to input data and to choose different outputs easily.

6. References

Anderson, C., Norberg, L. (1994), 'A method for variance estimation of non-linear function of totals in surveys, theory and software implementation', *Journal of Official Statistics,* Vol. 10.

Bellhouse, D. R. (1985), 'Computing methods for variance estimation in complex surveys', *Journal of Official Statistics,* Vol. 1.

Bethel, J. (1989), 'Sample allocation in multivariate surveys', *Survey Methodology*, 15, pp. 47–57.

Deville, J. C., Särndal, C. E. (1992), 'Calibration estimators in survey sampling', Journal of the American Statistical Association, Vol. 87, pp. 367–382.

Estevao, V., Hidiroglou, M. A. and Särndal, C. E. (1995), 'Methodological principles for a generalised estimation system at statistics Canada', *Journal of Official Statistics*, Vol. 11, N.2, pp.181–204.

Falorsi, P. D., Ballin, M., De Vitiis, C. and Scepi, G. (1998), 'Principi e Metodi del Software Generalizzato per la Definizione del Disegno di Campionamento nelle Indagini sulle Imprese Condotte dall'Istat', *Statistica Applicata*, Vol. 10, N.2.

Falorsi, S. and Rinaldelli, C. (1998), 'Un Software Generalizzato per il Calcolo delle Stime e degli Errori di di Campionamento', *Statistica Applicata*, Vol. 10, N.2.

Särndal, C. E., Swensson, B. and Wretman, J. (1992), *Model assisted survey sampling*,. Springer-Verlag, New-York.

Schnell, D., Kennedy, W. K., Sullivan, G., Park, K. P. and Fuller, W. A. (1988), 'Personal computer variance software for complex surveys', *Survey Methodology*, Vol. 14, pp. 59–69.

Shah, B. W., Lavange, L. M., Barnwell, B. J., Killinger, K. E. and Wheeless, S. C. (1989), 'SUDAAN, procedures for descriptive statistics user's guide', *Research Triangle Institute Report*.

Singh, A. C. and Mohl, C. A. (1996), 'Understanding calibration estimators in survey sampling', *Survey Methodology*, Vol. 22, N.2, pp. 107–115.

Verma, V. (1995), 'Weighting for Wave 1', *Eurostat documentation*, doc. PAN 36/95.

Statistical research in the fifth framework programme: an update for 2000

1. Introduction and background

The purpose of this paper is to present a short update of activities in the fifth framework research programme in official statistics since the beginning of the year 2000. We will assume that the reader is familiar with the basic concepts and terms of the research programme and they will not be repeated here. For an introduction and terminology, see Wickens (1999) in a previous issue of this journal. For online access, the VIROS web site (**http://europa.eu.int/comm/eurostat/research/**) contains much additional and constantly updated information. In particular, for previous programmes, see the documents under the 'DOSES', 'DOSIS', 'SUP.COM' links on the VIROS page. For the present programme, see the information under the 'R & D programmes' and 'What's new' links in the VIROS page.

We will cover the progress by the central research themes, tools and methods, statistical indicators of the new economy (SINE) and technology transfer. The research projects were accepted under different calls for proposals; for reference purposes, the calls are also mentioned.

The research themes cover an extremely large area in official statistics and we will be able to point out merely the main features of the programme. As will be seen, many new projects are starting or have been received during the year. Due to lack of space, only a brief list of the projects will be given, with pointers to additional information on the web. Only the projects monitored by Eurostat are mentioned.

2. Statistical tools and methods

2.1. Projects resulting from the first call

Between January and March 2000, the successful projects which were received from the first call in 1999 were started by the research teams. The following table gives a crude first level classification of the research areas involved.

Classification	Project acronym
1. Methodological issues	IPIS, Chintex, Clamour
2. Advanced technologies for data collection	IQML
3. Quality issues	Euredit
4. Data analysis and statistical modelling	SPIN!, APPETISE, IMPACT, BUSY
5. Multiple data sources, integration	Metaware, Mission
6. Dissemination and disclosure control	
7. Other innovative applications of information technologies	VL-CATS, X-Statis

The VIROS web site (VIROS → 'R & D programmes' → 'First call') gives detailed descriptions of the projects. Most projects have links to their web sites for additional information and the project results.

2.2. Third call projects

Additional projects on tools and methods will start at the beginning of the year 2001. The following projects will complement the previous table.

Classification	Project acronym
1. Methodological issues	
2. Advanced technologies for data collection	Mantle (second call), FLASH
3. Quality issues	Dacseis, Eurarea
4. Data analysis and statistical modelling	ASSO
5. Multiple data sources, integration	Metanet, Inspector
6. Dissemination and disclosure control	CASC
7. Other innovative applications of information technologies	Vitamin-S, Statlas

Additional information on the project will appear soon on the web site (VIROS → 'R & D programmes' → 'Third call').

3. Research themes in Tools and Methods

The research programme has now reached a stage where results and experiences from earlier projects lead to an accumulation of know-how. Simultaneously, research on specific

themes is no longer dependent on a single project but integration of work from several projects is feasible. Two examples, metadata and data quality, are given below.

3.1. Metadata

In the fourth framework programme, several projects already had a strong statistical metadata content and this has continued in the fifth programme. Projects with a metadata component such as IMIM, Idaresa, Addsia, Faster, IPIS, IQML, Mission, Metaware (see VIROS for details) belong to this group. At the same time, several international bodies are dealing with the harmonisation of metadata; a need for convergence in the metadata area has been identified.

With the critical mass coming from these projects, a new type of project, a thematic network and cluster called Metanet, is now starting to develop and integrate proposals for views and common standards and to disseminate the results.

3.2. Data quality and best practices

Again, several projects have been addressing the issues of data quality, specifically the areas of imputation, editing and (variance) estimation. The projects Autimp (fourth framework), Euredit, Eurarea and Dacseis address new and innovative aspects of these classical practices in official statistics. A common approach is methods comparison where the performance of existing and new methods are compared empirically, on real datasets or in simulated but realistic populations. Best practices and methods comparison are also addressed in the area of statistical confidentiality (CASC), in harmonisation (Chintex) and various other fields. It is foreseen that the projects will both cluster among themselves and contribute to general 'best practices' guidelines in the data quality area.

Several other themes such as data capture, data integration, time-series techniques, automated coding, EDI, etc., arise in a similar way from the present programme. It is expected that in the forthcoming years many of these themes will reach a critical mass and through consolidation will lead to a considerable contribution to the development of the European statistical system.

3.3. SINE

In the second call, a number of projects were received and have started in the last months of the year 2000. The projects specifically addressing SINE are Eicstes, STING, and Newkind, creating indicators of the new economy by exploring the web, patent information and 'knowledge-basis' of economy, respectively. Additional information on the project will appear soon on the web site (VIROS _ 'R & D programmes' _ 'Second call').

The indicator part of the programme is relatively new and at present quite fragmentary. While SINE is a key area to be promoted in the programme, it is too early to evaluate its suc-

cess or the themes emerging from it. It is clear that the conceptual/theoretical dimension is far from being addressed adequately and that there is an urgent need to implement this research area to statistics production.

A new call, restricted to SINE and with special emphasis on the *e*Europe initiative was launched in October 2000, with a deadline in January 2001 and a further call will open in early 2001. It is hoped that the proposals received in these calls will result in a consolidation and usefulness we have seen previously in the tools and methods sector of the programme.

4. Technology transfer

A specific type of research action, an accompanying measure, with the acronym Amrads was received in the third call. Amrads, starting in January 2001, has the ambitious goal to ensure that the research results, including the best practices, from the research programme are transferred to and taken up by the European statistical system. Amrads will base its activity on six focal thematic networks:

- statistical disclosure control;
- advanced technologies for data collection;
- quality issues;
- time-series analysis and seasonal adjustment;
- business registers and administrative sources;
- multi-source data integration.

5. Future prospects

The present EPROS programme will continue in the annual work programmes of the IST programme. In the work programme 2001, research in official statistics will fall under Cross-Programme Action 7 and a call for proposals is foreseen in January 2001 for both tools and methods and indicators.

In the year 2000, a first step to plan future research beyond the present fifth framework programme was taken. Eurostat launched the planning of the statistical content of the sixth framework programme, starting in 2003, by inviting a group of international experts to produce a research programme in official statistics. A first vision paper was introduced in the EPROS meeting in December 2000 to the representatives of the Member States. With several iterations and comments from the national statistical institutes and the scientific community, the group is expected to produce a document forming the basis of research in the first half of the year 2001. On the 21 February 2001, the Commission proposed a new framework programme. This is aimed to reinforce statistical research.

6. Further information

Further information on statistical research activities past, present and future within the Framework Programmes of the European Union can be obtained from Eurostat: please contact Mr Jean-Louis Mercy. Phone: (352) 4301-34862; fax (352) 4301-34149; e-mail: Jean-Louis.Mercy@cec.eu.int.

At the time of going to press, a publication with the release date of end April 2001 was being finalised. This contains a synopsis of projects currently being carried out within the fifth framework programme. Copies of the publication could be obtained by contacting the person named above.

Additional information could be obtained from the European Commission web site (*http://www.cordis.lu/ist*) and from the Eurostat research web site **(***http://europa.eu.int/comm/eurostat/research*).

7. References

Wickens, J. (1999), 'Research and development in (official) statistics: fifth RTD programme 1999–2002', *Research in Official Statistics*, 1, Vol. 2, pp. 91–98.

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