Chapter 13

Software Demonstrations

■ MatchWare Product Overview

Matthew A. Jaro, MatchWare Technologies, Inc.

Probabilistic linkage technology makes it feasible to link large data files and achieve results governed by mathematical principles which adhere to statistically valid standards. The problem addressed by this methodology is that of matching two data files under conditions of uncertainty. The objective is to identify and link records which represent a common entity whether that entity is an individual, a family, an event, a business, an institution, or an address. As an alternative the goal might be to unduplicate a single data file or to group records by categories of commonality. Each field participating in the linkage comparison is subject to error which is measured by the probability that the field agrees given a record pair matches versus the probability of chance agreement of its values. Thus, when one calculates the likelihood of a correct match or link while allowing for incomplete and/or error conditions within the records, the process is said to be probabilistic. I. P. Fellegi and A. B. Sunter pioneered record linkage theory in the late 1950s. The first practical implementation of probabilistic linkage methodology in the United States was originally designed, programmed, and tested by Matt Jaro on behalf of the U. S. Census Bureau in 1985, while conducting research into establishing a model to support census coverage undercount evaluation and analysis.

Probabilistic record linkage methodology is imperative if computers are to consistently and effectively replicate the evaluation and judgment process of human clerks attempting to link common records. The ideal goal is to have the computer emulate the intuitive thought process of a human being as they might review, judge, evaluate, measure, and score linkage qualifications of records representing commonality.

MatchWare's development, systems design, and programming staff rigorously and strictly adhere to ANSI-C programming language standards for all software implementations. As a result, MatchWare software has achieved an exceptional level of cross-platform portability and can be integrated into a wide range of application solution specific systems. Following are the products currently offered by the company:

> AutoStan is an intelligent pattern recognition parsing system which conditions records into a normalized/standardized fix fielded format. AutoStan optimizes the performance of any linkage or matching system which utilizes consumer or business names and/or address data as identifiers during a match comparison. AutoMatch is a state-of-the-art software implementation of probabilistic record linkage methodology for matching records under conditions of uncertainty. AutoMatch simulates the thought process a human being might follow while examining and identifying data records representing a common entity or event. AutoMatch's comparative algorithms manage a comprehensive range of data anomalies and utilize frequency analysis methodology to precisely discriminate weight score values.
>
> AutoStan and AutoMatch are stand-alone, self-contained software systems which include numerous support utilities and require no other ancillary software. Both systems are generalized and support a wide range of mission critical record linkage applications. AutoStan and AutoMatch adhere to widely accepted standards of statistical methodology to ensure valid results and the highest levels of data integrity. Users have ready access to Rule/Table Portfolios in order to calibrate the software for their particular requirements. MatchWare/CL is a callable library (API) version of AutoStan and AutoMatch functionality in executable module form. MatchWare/CL utilizes AutoStan and AutoMatch Rule/Table Portfolios, weight scoring formulae, and statistical algorithms. MatchWare/CL is compatible with any database manage-

ment system or user interface, and has been integrated into a variety of application solution specific systems.

Both AutoStan and AutoMatch are generalized and support a wide range of mission critical health data registry, geocoding, and database marketing applications.

For more information, contact Max Eveleth Jr., Executive Vice President, MatchWare Technologies, Inc., 153 Port Road - 2nd Floor, Kennebunk, ME 04043-5135; Phone: (207) 967-2225; Fax: (207) 967-8362; or e-mail: *meveleth@matchware.com* .

■ :- and J-ARGUS: Software Packages for Statistical Disclosure Control

Anco J. Hundepool, Agnes Wessels and Lars van Gemerden, Statistics Netherlands

In recent years, Statistics Netherlands has developed a prototype version of a software package, ARGUS, to protect microdata files against statistical disclosure. The launch of the SDC-project within the 4th framework of the European Union had enabled us to make a new start with the development of software for Statistical Disclosure Control. More information on the SDC-project can be found at *http://www. cbs.nl/sdc.*

This prototype has served as a starting point for the development of :-ARGUS, a software package for the SDC of microdata files. The aim is to produce a data file for which the risk of disclosure has been minimized and which can be supplied to researchers and other users. The basic principle of :-ARGUS is that frequency tables of combinations of identifying variables are inspected. If the frequency in a cell is too low, it means that a certain combination does not occur frequently enough in the population and that the corresponding records, therefore, can easily be identified by an intruder. Techniques used in :-ARGUS to solve these problems are global recoding (using less detailed code lists) and local suppression (imputing missing values in these combinations).

This SDC-project, however, also plans to develop τ-ARGUS -- software devoted to the SDC of tabular data. τ-ARGUS takes the dominance-rule as a starting point to identify the unsafe (primary) cells, although other rules could be used, as well. Global recoding is applied to reduce most of the unsafe cells and optimization techniques are used to find a optimal set of secondary cells, which must be suppressed to protect the primary unsafe cells.

Both :- and τ-ARGUS have been developed for Windows 95 PC's. However, we have developed ARGUS using Borland C++, which raises the possibility of easily generating modules (the parts of ARGUS accessing large datafiles) to be used on other platforms like UNIX.

Further information can be obtained from Anco Hundepool, Department for Statistical Methods, Statistics Netherlands, P.O. Box 4000,2270 J.M. Voorburg, The Netherlands; tel: +31-70-3375038; fax: +31-70-3375990; or e-mail: *argus@cbs.nlofahnl@cbs.nl.*

■ OX-LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version

Leicester E. Gill, University of Oxford, UK

The micro-computer version of OX-LINK is being used to match a dataset containing 150,000 hospital discharge and vital records. The matching and linking process is undertaken in three stages:

- The creation of an ONCA header, which is attached to every record on the dataset.
- Sorting the file on the keys which are stored in the ONCA header.
- Running OX-LINK to create a file of potential match pairs. A number of output files are produced which are used for verification of the match by clerical staff. The threshold weight matrix can be edited using Microsoft EDIT, and the whole of this stage can be rerun to demonstrate the changes in acceptance weight.

For more information, write to:

L. E. Gill
University of Oxford
Unit of Health-Care Epidemiology
Institute of Health Sciences,
Old Road, Headington, Oxford, OX37LF

or e-mail: *leicester.gill@clinical-epidemiology.ox.ac.uk* or *lester@pgme.warwick.ac.uk* .

■ Software for Record Linkage of Primary Care Data

John R. H. Charlton, Office of National Statistics, UK

The UK Royal College of General Practitioners collected data on all consultations in sixty practices in England and Wales over a one-year period 1991/92. In addition, socio-economic data were collected by survey from all patients registered with these practices. Each practice was sent a copy of its own data and the data from all the practices were combined into one dataset containing information on about 1.5 million consultations and about half a million patients.

The software demonstrated was written so that individual practices could easily access their own data, without specialised database software, or knowledge of the data structures and codes. Later, a modified program was written so that the Royal College of General Practitioners could extract data from the combined data from all practices. An anonymized version of the dataset was made available to other researchers and a further modified version of the program was produced for use with this dataset.

The program has two main functions. Firstly, to enable researchers to link different parts of the dataset, particularly patients and diseases, and secondly, to provide data summaries such as frequencies and rates. It is based on the Paradox database software and written in PAL, the language provided with Paradox for DOS. An installation program is provided to convert the ASCII files provided into the Paradox tables used by the program. The program can be run under either DOS or Windows.

For more information, contact Judith Charlton, 195 Warren Road, Orpington, Kent, BR6 6ES, U.K.; e-mail: *100025.1356@compuserve.com* .

# GRLS -- Record Linkage

Kathy Zilahi, Statistics Canada

This product addresses the problem of trying to link records where no unique identifiers exist. Our Generalized Record Linkage System (GRLS) was developed to enable such problem linkages to be successfully accomplished. GRLS improves both the quality and the ease of your linkage.

Features

Based on statistical decision theory, GRLS breaks a linkage operation into three steps:

- Search: Using comparison rules and associated linkage weights, the files are matched and a database of potential links is created.
- Decide: Linkage weights are refined and by using threshold weights, the potential links are divided into sets of possible and definite links.
- Group: Records which pertain to the same entity (person, business, etc.) are grouped together (the output of GRLS).

The GRLS record linkage system:

- provides a convenient framework for testing linkage parameters;
- allows concurrent users for each linkage project;
- allows background or interactive linkage;
- eliminates confusion (and paper!) with on-line help;
- makes your final linkage fast, cheap and accurate.

Applicability

GRLS handles one-file (internal) and two-file linkages such as:

- unduplicating mailing address lists (one-file);
- bringing hospital admission records together to build "case histories" (one-file);
- epidemiology studies: e.g., linking a file of workers exposed to potential health hazards, to a mortality database for the purpose of detecting health risks associated with particular occupations (two-file).

Platform Specifications

GRLS uses a client-server architecture, where a PC is the client and a UNIX box is the server. The ORACLE relational database management system Version 7.3 with SQL*PLUS, PL*SQL, PRO/C, FORMS 4.5 runtime, GRAPHICS 2.5 runtime and a "C" compiler are also required. With ORACLE Version 7.3, distributed processing can easily be achieved by using either a remote or local host from a mainframe, mid-range computer, or PC.

Contact Information

For more information, contact Ted Hill, by phone: (613) 951-2394; fax: (613) 951-0607; or e-mail: *ted-hill@statcan.ca*; or Bonnie Rideout, by phone: (613) 951-1714; fax: (613) 951-0607; or e-mail: *bburges@statcan.ca* .