Web Systems That Disseminate Information But Protect Confidential Data

Alan F. Karr and Ashish P. Sanil
National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709–4006 USA
{karr, ashish}niss.org

1 Introduction

Statistical agencies have longstanding concern over confidentiality of their data — both identities of data subjects and sensitive attributes in the data (FCSM, 1994; Willenborg & de Waal, 1996). But agencies must also report information to the public. This tension between confidentiality and dissemination of statistical information (Duncan, *et al.*, 1993) is heightened by the emergence of the World Wide Web as a means of communication.

On the one hand, confidentiality is threatened by advances in information technology (IT), such as powerful capabilities for record linkage across multiple databases. Other new technologies, however, not only protect confidentiality, but also meet user needs in innovative ways. Here we describe Web–based query systems being developed by the National Institute of Statistical Sciences (NISS) that use IT to disseminate, in the form of statistical analyses, information derived from confidential databases, but protect confidentiality of the data (NISS, 2001).

Two such systems are described. The first (§2) uses geographical aggregation to disseminate survey data collected by the National Agricultural Statistics Service (NASS) at the highest resolution consistent with the risk criteria. *Table servers* (§3) disseminate marginal subtables of a large contingency table and feature dynamic assessment of disclosure risk, in light of previous queries, which also allows data to be probed most deeply in regions of user community interest.

2 The NASS System: Geographical Aggregation

The NASS database contains 194,410 records, from 30,500 farms, detailing use of 322 chemicals on 67 crops in the years 1996–1998. Record attributes are Farm ID, size in acres, crop, chemical, pounds of the chemical applied, state, county and year.

System Design and Disclosure Criteria. User queries are for *application rates* (pounds applied per acre) of certain chemicals on particular crops, ideally at the county level. (Currently, NASS releases data only at the state level.)

Disclosability of the application rate in a geographical unit is based on the (N, p)–rule (Willenborg & de Waal, 1996): the unit must contain at least $N = 3$ surveyed farms for the specified chemical, crop and year, and no farm surveyed in the unit may comprise more than $p = 60\%$ of the total acreage. At the county level, however, these rules do not work. More than one-half of counties are undisclosable, and a system that simply refused to answer such queries would lead to unacceptable user frustration.

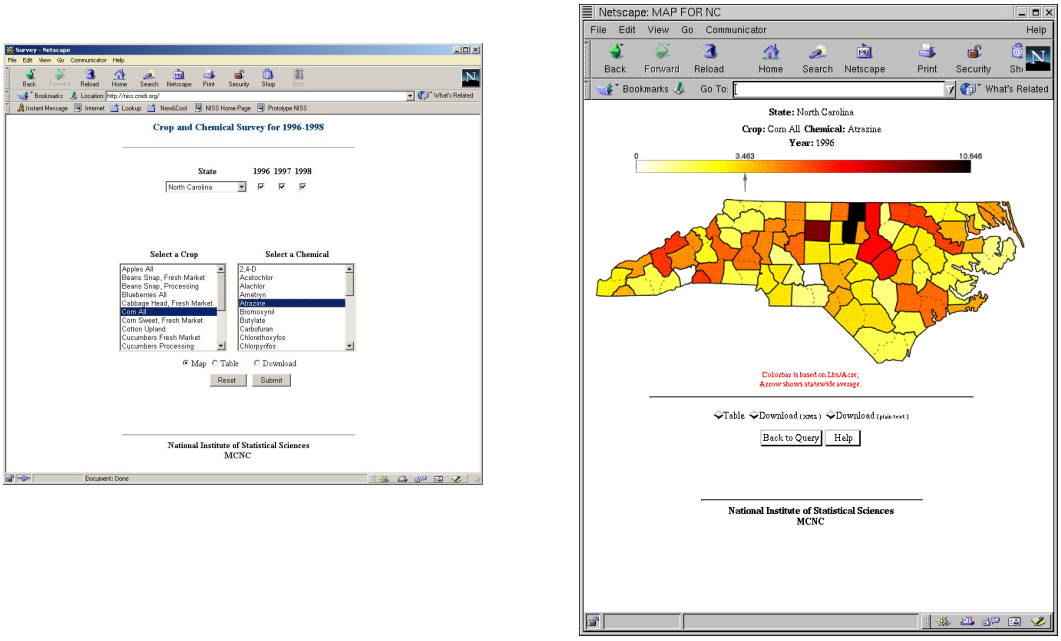

Figure 1: The NASS prototype. *Left:* Input screen, on which users select a state, year(s), crop and chemical. *Right:* Output screen with map displaying the requested application rate. Supercounties are colored according to the application rate of the chosen chemical on the chosen crop.

Aggregation Algorithms. The system built by NISS for NASS (Karr, *et al.*, 2001) uses geographic aggregation to achieve disclosability: undisclosable counties are merged with neighboring counties (in the same state) to form disclosable "supercounties."

Two heuristic algorithms have been developed (Karr, *et al.*, 2000), which share a common structure. Each examines the undisclosable (super)counties in a random order and merges them with neighboring (super)counties until only disclosable (super)counties remain. The *pure* algorithm favors leaving disclosable counties unmerged, preserving "purity" of their data. The *small* algorithm favors forming small supercounties by merging an undisclosable region with a neighboring region most likely to achieve disclosability.

Both algorithms randomize the order in which candidate mergers are considered, and break ties either randomly or on the basis of similarity of application rates in the merger candidates.

Implementation. Because the algorithms may produce supercounties that can be decomposed (Karr, *et al.*, 2000), the system employs a two–step procedure. First the small algorithm is run on a state as whole, and then the pure algorithm is run within each supercounty produced by small.

Figure 1 shows screen shots from the prototype system, which is accessible at niss.cnidr.org. The user selects a state and year(s) (screen not shown), then a crop and a chemical and finally an output format — map (shown in Figure 1), on-screen table or XML download.

Statistical Consequences of Aggregation. Like other disclosure limitation methods, aggregation alters statistical properties of the released information. Bayesian simulation approaches (Lee, *et al.*, 2001) permit analysis of aggregated data, as well as specification of additional informa-

tion (such as certain sample sizes) that must be released in order to enable the user to perform meaningful analyses.

3 Table Servers

The target database of a table server is a large (e.g., 40 dimensions with 4 categories each) contingency table, containing counts or sums. The essential characteristic of the table servers being developed by NISS is that they are *dynamic*: user queries for marginal sub-tables of the full table are assessed for disclosure risk in light of previously answered queries. Potential responses include the requested table (as text, visualized or in XML), its "projection" onto the released frontier (see below), risk-reduced modifications of the requested table and refusal of the query.

System Design. The central challenges to building such a system are development of the abstractions that underlie it and of scalable algorithms that implement the abstractions.

The query space $\mathcal{Q}$ is partially ordered by set inclusion (of variables in a table). The set $\mathcal{R}(t)$ of all tables released through time t contains both direct (in response to queries) and indirect (children of direct) releases, whose *released frontier* consists of the maximal elements of $\mathcal{R}(t)$. At t there are also an *unreleasable set* $\mathcal{U}(t)$ of subtables whose release would be too risky (see below), with an *unreleasable frontier* of minimal elements.

Underlying release decisions is a *risk criterion* **RC**: a requested but unreleased table T cannot be released if

$$\mathbf{RC}(\mathcal{R}(t) \cup T) > \alpha$$

where α is a threshold set by the system operators. Possible risk criteria include: (1) Accuracy of bounds calculated from $\mathcal{R}(t)$, especially for sensitive cells (Dobra & Fienberg, 2000); (2) Accuracy of reconstruction of the full table by iterative proportional fitting; (3) Predictive capability for sensitive variables in the full table; and (4) The number of tables whose marginals coincide with all elements of $\mathcal{R}(t)$ (or some other measure of the size of the set of all such tables).

Release rules select which requests for unreleased tables will be fulfilled, and may operate in real time or in batch mode and off–line. Release decisions may take into account (1) The risk criterion; (2) Which tables T' become too risky to release as a consequence of releasing a requested table T; and (3) The *value* of releasing T.

Prototype System. A prototype table server, written as a Java application, is shown in Figure 2. It operates on an 8–dimensional table of data from the Current Population Survey. Figure 2 also shows (system operators) the potential effect of releasing a particular 5–way subtable.

References

Dobra, A., and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Nat. Acad. Sci.* **97(22)** 11885–11892

Duncan, G. T., de Wolf, V. A., Jabine, T. B., and Straf, M. L. (1993). Report of the panel on confidentiality and data access. *J. Official Statist.* **9** 271–274.

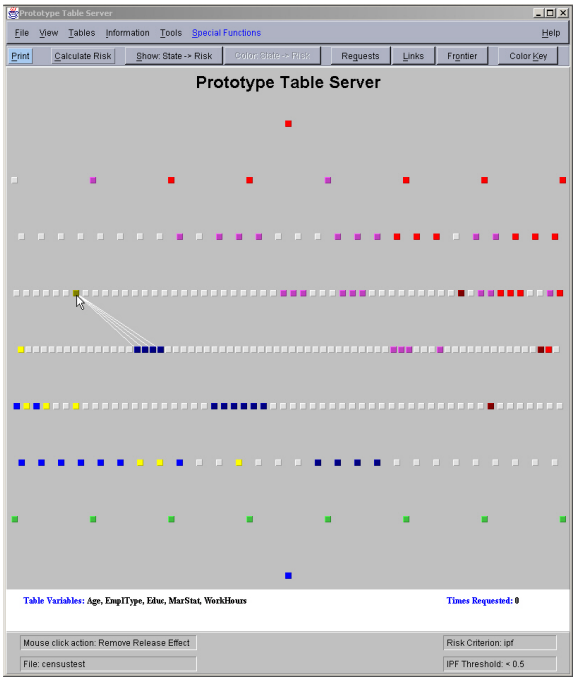

Figure 2: Prototype table server. Shown are a visualization of the query space, core releases at the start of system operation (green), direct releases (yellow), indirect releases (blue), unacceptably risky releases (red) and the potential effect (dark blue, magenta and dark red) of releasing the 5–way table over which the cursor is positioned.

Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology*.

Karr, A. F., Lee, J., Sanil, A., Hernandez, J., Karimi, S., and Litwin, K. (2000). Web-based systems that disseminate information from data but preserve confidentiality. Technical Report, NISS. Available on line at www.niss.org/dg/technicalreports.

Karr, A. F., Lee, J., Sanil, A., Hernandez, J., Karimi, S., and Litwin, K. (2001). Disseminating information but protecting confidentiality. *IEEE Computer* **34(2)** 36–37.

Lee, J., Holloman, C., Karr, A. F., and Sanil, A. P. (2001). Analysis of aggregated data in survey sampling with application to fertilizer/pesticide usage surveys. Technical Report, NISS. Available on line at www.niss.org/dg/technicalreports.

National Institute of Statistical Sciences (2001). Digital Government Project Web site. Available on-line at www.niss.org/dg.

Willenborg, L., and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. Springer–Verlag, New York.