NISS

Disseminating Information but Protecting Confidentiality

Alan F. Karr, Jaeyong Lee, Ashish Sanil,
Joel Hernandez, Sousan Karimi,
and Karen Litwin

Technical Report Number 107
October, 2000

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Disseminating Information but Protecting Confidentiality

Alan F. Karr, Jaeyong Lee, Ashish Sanil
National Institute of Statistical Sciences, Research Triangle Park, NC
Joel Hernandez, Sousan Karimi, Karen Litwin
MCNC, Research Triangle Park, NC

Federal statistical agencies have longstanding concern over confidentiality of their data (sample surveys and censuses). Both the identities of data subjects and sensitive attributes in the data must be protected [1, 2]. But the agencies also have an obligation to report information to the public.

This tension between confidentiality and dissemination of statistical information [3] arises equally sharply in non-governmental contexts, from electronic medical records [4] to E-commerce transaction data.

Confidentiality is threatened by advances in information technology, 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 a system (see `www.niss.org/dg`) being developed by the National Institute of Statistical Sciences for the National Agricultural Statistics Service (NASS), which disseminates survey data on usage of agricultural chemicals in far greater geographical detail than previously, but protects the identities of farms in the survey.

The NASS Data. The database contains 194,410 records, from 30,500 farms, detailing use of 322 chemicals (fertilizers, fungicides, herbicides, pesticides) 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.

User queries are for *application rates* (pounds applied per acre) of certain chemicals on particular crops, ideally at the county level. Currently NASS releases application rates only at the state level. The system we describe produces more informative aggregations than state-level releases, but preserves confidentiality.

Aggregation for Disclosure-Risk Reduction. For the application rate in a geographical unit to be disclosable, NASS requires that two widely employed rules [2] be satisfied. The *N-rule* requires that the unit contain at least $N = 3$ surveyed farms for the specified chemical, crop and year. The *p-rule* prohibits a dominant farm that comprises more than $p = 60\%$ of the total acreage of all farms surveyed in the unit.

At the county level these rules do not work: more than 50% of counties are undisclosable. Our system aggregates undisclosable counties with neighboring counties (in the same state) to form disclosable "supercounties," allowing NASS to release data at the highest resolution consistent with the risk criteria.

Aggregations must be computed automatically in response to user queries. Computing aggregations can be formulated as a (NP-hard) combinatorial optimization problem over the edge-set of the adjacency graph of the counties in a state, and solved (which we have done) using simulated annealing methods, but long running times make this infeasible in practice. Instead, we employ heuristic, "greedy" algorithms [5], which produce aggregations differing only insignificantly from those produced by simulated annealing.

Two heuristic algorithms have been developed, which share a common structure: examine the undisclosable (super)counties in a random order and merge them with a neighboring (super)county until only disclosable (super)counties remain.

The algorithms differ only in the rule that governs merging. The pure rule favors leaving disclosable counties unmerged (preserving "purity" of their data), but can create large supercounties comprised of many undisclosable counties. The small rule, by contrast, 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 randomly). Each can produce aggregations in which some supercounties can be decomposed [5]. To alleviate this, our implementation first runs the small algorithm, and then runs the pure algorithm within each supercounty produced by small. This composite procedure works fast and well.

NASS System Architecture and Operation. Figure 1 shows two screenshots from the prototype NASS system, which is accessible at `niss.cnidr.org`.

The user first selects (left panel of Figure 1) a state and year(s) of interest. JavaScipt routines then dynamically generate drop-down menus of relevant crops and chemicals. The user next selects either a crop (in which case the chemical menu is regenerated to contain only chemicals applied to that crop) or a chemical (causing the crop menu to be regenerated). Finally, an output format is selected: map (the default), on-screen table or XML download. The XML DTD mirrors the hierarchical nature of the aggregated data.

If not available from a previous query, the aggregation is computed on-the-fly. The result is stored in case the query is received again, and transaction information is written to a query history database.

Map output is shown in the right panel of Figure 1. Supercounties are colored according to the application rate of the chosen chemical on the chosen crop; the color bar also shows the state-wide average rate. Supercounty and county—within–supercounty boundaries are shown, but differently. Multiple years appear on separate maps with a common color scale.

Concluding Comments. Citizen access to data and information is an essential responsibility of the Federal government. The system described here is a step toward using the Web to meet that responsibility.

More complex databases, queries and privacy concerns lead to additional challenges. These include queries whose disclosure risk depends on which queries have been answered previously, risk computation and reduction for queries entailing integration of multiple databases, and problem formulations and tools that allow agencies to balance the value to society of releasing information against disclosure risk.

References

[1] Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology*, May 1994.

[2] L. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. Springer-Verlag, New York, 1996.

[3] G. T. Duncan, V. A. de Wolf, T. B. Jabine, and M. L. Straf. Report of the Panel on Confidentiality and Data Access. *Journal of Official Statistics*, 9:271–274, 1993.

[4] Who knows your medical secrets? *Consumer Reports*, pages 23–26, August 2000.

[5] A. Karr, J. Lee, A. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Web-Based Systems that Disseminate Information from Data but Protect Confidentiality. Technical report, National Institute of Statistical Sciences. Available at `http://www.niss.org/dg/technicalreports.html`, 2000.

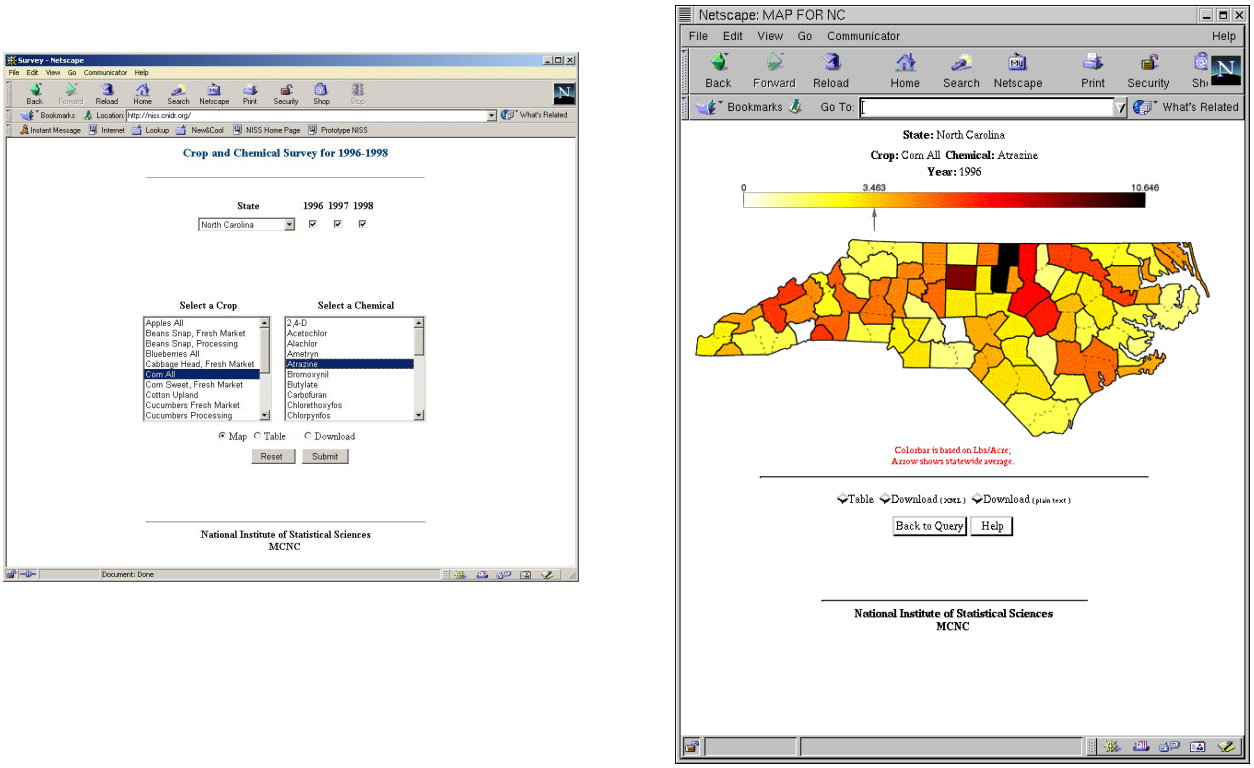

Figure 1: The NASS prototype. *Left:*Input screen, on which users a state, year(s), crop and chemical. *Right:* Output screen with map displaying the requested application rate (using artificial data).