

## Chapter 1

# Web-Based Systems that Disseminate Information from Data but Protect Confidentiality

Alan F. Karr, Jaeyong Lee, Ashish P. Sanil  
*National Institute of Statistical Sciences*  
*Research Triangle Park, NC, USA*

Joel Hernandez, Sousan Karimi, Karen Litwin  
*MCNC*  
*Research Triangle Park, NC, USA*

**Abstract:** The Internet provides an efficient mechanism for Federal agencies to distribute their data to the public. However, it is imperative that such data servers have built-in mechanisms to ensure that confidentiality of the data, and the privacy of individuals or establishments represented in the data, are not violated. We describe a prototype dissemination system developed for the National Agricultural Statistics Service that uses aggregation of adjacent geographical units as a confidentiality-preserving technique. We also outline a Bayesian approach to statistical analysis of the aggregated data.

**Key words:** Data confidentiality, privacy, Web dissemination, digital government, aggregation, statistical analysis

## 1. INTRODUCTION

There has been longstanding concern in the United States and elsewhere over the confidentiality of statistical data, especially data gathered by the Federal government in sample surveys and censuses (see Appendix B for examples). Confidentiality may be mandated by law, prescribed by agency

practices or promised to respondents. Often, confidentiality must be preserved in order to ensure the quality of the data: respondents may lie if they believe that their privacy is threatened.

At the same time, government agencies have an obligation to report their data, or information derived from the data, and they recognize the need for some balance between strict confidentiality and the benefits derived from the release of statistical information (Duncan, *et al.*, 1993).

The Internet has emerged as a natural and efficient mode for disseminating Federal data (Schorr & Stolfo, 1997). However, the ease of access and potentially large quantities of data that Web-based systems can provide intensify the need for techniques that prevent disclosure of confidential data. Computing power and ubiquity of databases make record linkage (see Appendix A) and other means of breaking confidentiality surprisingly easy (Keller-McNulty & Unger, 1993). On the positive side, the opportunity exists to implement new technologies to protect confidentiality, which can also meet user needs in new ways.

Research on this topic is being carried out at the National Institute of Statistical Sciences (NISS), working with collaborators at Carnegie Mellon University, Los Alamos National Laboratory, the Ohio State University and MCNC under an NSF-sponsored Digital Government project. The project Web site is [www.niss.org/dg](http://www.niss.org/dg). Several aspects of the problem make this a challenging cross-disciplinary effort: devising strategies, some that reflect the entire history of queries to a system, to evaluate and reduce disclosure risk; and exploring issues of system design, user interfaces, scalable data structures and new Internet technologies.

In this article, we describe a prototype system developed for the National Agricultural Statistics Service (NASS). The system disseminates survey data concerning on-farm usage of chemicals (fertilizers, fungicides, herbicides and pesticides) in far greater geographical detail than previously, but protects the identities of farms in the survey. Confidentiality is preserved by means of geographical aggregation: data from adjacent counties are aggregated to the level of disclosable "supercounties."

## 2. THE NASS DATA

The data consist of on-farm use of agricultural chemicals on various crops in 1996-1998, collected by NASS through an annual survey of farms.

The database contains 194,410 records collected from 30,500 farms, with information on the rates of usage of 322 chemicals on 67 crops (field crops, fruits and vegetables). For our purposes, each data record can be thought of as containing Farm ID, state, county, year, farm size in acres, crop,

chemical, and the number of pounds of the chemical applied to that crop. (The real database is more complex, involving quantities such as number of applications and adjustment weights.)

User queries to the system are for *application rates* (pounds applied per acre) of certain chemicals on particular crops in geographical regions of interest. Ideally, information would be released at the county level. Currently, however, because of confidentiality concerns, NASS releases application rates only at the state level.

### 3. AGGREGATION FOR DISCLOSURE RISK REDUCTION

The confidentiality concern of NASS is to protect the identities of farms in the survey. Information cannot be disclosed that would enable a user to estimate accurately the chemical usage on a particular farm (Dalenius, 1977). Such a disclosure would breach the respondent confidentiality promised by NASS. (Adjustment weights are also confidential, for technical statistical reasons.)

For the application rate in a geographical unit to be disclosable, NASS requires that two widely employed rules (FCSM, 1994; Willenborg & de Waal, 1987) 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.

The underlying rationale for these rules, which collectively we term the  $(N, p)$ -rule, is that farms in a sample containing too few farms or a farm whose size dominates a sample are susceptible to both identity and attribute disclosure risks (see Appendix A).

At the county level the  $(N, p)$ -rule with  $N = 3$  and  $p = 60\%$  does not work: more than one-half of counties are undisclosable. Simply refusing to answer such queries would lead to unacceptable user frustration. Instead, the NISS 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, since there too many (State, year, crop, chemical) combinations to permit manual aggregation on a case-by-case basis. We employ two "greedy" algorithms (see Appendix C) based on the following heuristic procedure: Examine the undisclosable (super)counties in a random order and merge them with a neighboring (super)county according to some criterion for desirability of merging. Continue until only disclosable (super)counties remain.

The algorithms differ only in the rule that governs the merging process. The **pure** rule directs the merging of counties in a manner that favors leaving the disclosable counties alone, thereby preserving the “purity” of their data. Instead, it merges the undisclosable counties among themselves insofar as possible. This procedure does preserve purity of as many disclosable counties as possible, but it can create large supercounties comprised of many undisclosable counties.

The **small** rule, by contrast, favors forming small supercounties by merging an undisclosable region with the neighboring region most likely to achieve disclosability. In practice, judging by visual inspection, the **small** rule produces satisfactory aggregations.

Both algorithms randomize the order in which candidate mergers are considered, breaking ties either randomly or on the basis of similarity of application rates. Either can produce aggregations in which some supercounties may be decomposed into smaller but nevertheless disclosable supercounties (see Appendix C). To alleviate this, we employ a two-step process. First, the **small** algorithm is run to produce an initial aggregation, and then the **pure** algorithm is run within each supercounty produced by **small**, in order to determine if it can be decomposed.

This composite procedure works remarkably well. It is also very fast: test runs for aggregation of states containing 100 counties take less than a millisecond per run, even on a 233 MHz Pentium PC running Linux.

Moreover, it produces aggregations that are as good as those produced by complex, formal optimization procedures. The aggregation problem can be formulated as a (NP-hard) combinatorial optimization problem over the edge-set of the adjacency graph of the counties in a state. One advantage of doing this is that the optimization framework allows explicit incorporation of “goodness” of aggregations, allowing preference, for instance, for aggregating counties that lie within the same watershed. The combinatorial optimization problems can be solved using computationally intensive simulated annealing methods, but long running times make this infeasible in practice. In test cases where we have used both methods, there is no significant difference between the characteristics of the aggregations they produce, and we conclude that the heuristic methods are adequate for the application we describe here.

#### 4. NASS SYSTEM ARCHITECTURE AND OPERATION

The prototype NASS system (Karr, *et al.*, 2001) is accessible at [niss.cnidr.org](http://niss.cnidr.org). Figure 1 shows schematically the system architecture. The

system has been implemented on a Sun SPARCSTATION 10 running Solaris 2.6. The NASS survey database and the Query History Database (QHDB) are maintained by an Oracle 8i relational DBMS. An Apache HTTP server forms the front end, allowing users to interact with the system via a Web browser.



Figure 1. NASS System Architecture

As shown in Figure 2, the user first selects from an HTML form a State containing the counties for which chemical usage information is desired. Once a State is selected, CGI scripts written in Perl query the database and retrieve lists of crops grown in the State and chemicals used there. JavaScript routines then generate drop-down menus that let the user select among only the crops grown and chemicals used in the selected State.

Next, the user either selects a crop, in which case the Perl/JavaScript routines regenerate the menu for the chemicals, showing only those used on the selected crop, or the user selects a chemical, which leads to an updated menu of crops. At this point (or earlier), one or more years may be selected. The user may also override the default map-based output and specify another output format before submitting the query.



The directory and file naming convention we use makes it easy to identify and retrieve results corresponding to previous queries.

The default mode for reporting query results to the user is a map, as shown in Figure 3. Supercounties are colored according to the application rate of the chosen chemical on the chosen crop, and 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, but share a common color scale. The aggregations may differ year-to-year for a multiple-year query; this minimizes the degree of aggregation. At the expense of (much) greater aggregation, the system could force the same aggregation for multiple years.

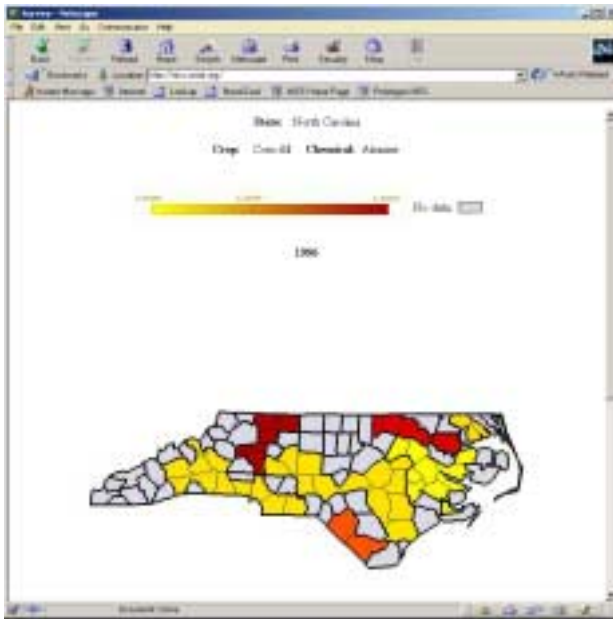


Figure 3. Map Output Screen

The user also has the option of viewing in tabular form the numerical values of the application rates underlying the map. Finally, the aggregated data can be downloaded in XML format, using a DTD that reflects the hierarchical nature of data aggregated at the supercounty level.

Navigation aids enable the user easily to modify the current query, or to request different forms of output. Navigation and query updates are facilitated by the storing the current session parameters in a hidden frame on the user's browser. Also, the system maintains internal status flags so that abnormal exits and disruptions can be handled appropriately.

## 5. STATISTICAL IMPLICATIONS OF AGGREGATION

While aggregation prevents disclosure, it may also distort the data. Thus, from a statistician's point of view, a question that immediately arises is: How should one go about analyzing the aggregated data in order to make informative inferences about the surveyed population?

In this section, we present a sketch of a Bayesian simulation approach for the analysis of such aggregated data. This methodology is the focus of Lee, *et al.* (2001), to which the reader is referred for details.

We abstract the NASS system as follows. Let  $\mathbf{P} = \{Y_1, \dots, Y_N\}$  be the data for the population (of farms) of interest, where  $N$  is the size of the population. Initially, we suppose that farms have only one attribute. Suppose that the disseminator samples  $\mathbf{S} = \{y_1, \dots, y_n\}$ , comprising the surveyed farms, using simple random sampling from  $\mathbf{P}$ .

The disseminator also draws a partition  $\kappa = \{\kappa_1, \dots, \kappa_k\}$  of the index set  $\{1, 2, \dots, n\}$  from a distribution  $p(\kappa | \mathbf{S})$ , representing the units of aggregation. Often,  $p(\kappa | \mathbf{S})$  is statistically independent of  $\mathbf{S}$ . This occurs, for example, if  $\kappa_i$  is defined by geographical units such as counties. On the other hand, in the NASS setting,  $\kappa$  depends on sampled values  $y_i$ , which leads to the need for the Bayesian approach described here.

Let  $n_i$  be the number of observations in the  $i^{\text{th}}$  partition set  $\kappa_i$ , for  $i = 1, \dots, k$ . The disseminator aggregates the sample over the partition  $\kappa$  and releases  $\mathbf{A} = \{(\bar{y}_1, n_1), \dots, (\bar{y}_k, n_k)\}$ , where  $\bar{y}_i = (1/n_i) \sum_{j \in \kappa_i} y_j$ .

To illustrate analysis of aggregated data, consider estimation of the population variance. The usual estimator  $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$  of the population variance  $S^2 = \sum_{i=1}^N (Y_i - \bar{Y})^2 / (N-1)$  cannot be recovered from  $\mathbf{A}$  in general. However, a natural extension of the usual estimator in this case leads to the estimator

$$s_a^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2.$$

This estimator, though, is not guaranteed to be unbiased unless  $\kappa$  is independent of the sample  $\mathbf{S}$ .

Thus, the dependence of the partition on the data values of the sample renders the usual estimation procedures deficient. Moreover, to keep this description simple, we have taken the data  $Y_i$  to be simple real-valued random variables; whereas in the NASS system they are ratios of pounds

applied to acres. The analysis of ratios introduces additional complexities. It is clear that any methodology that addresses the aggregated data problem must not only accommodate the dependency between  $\kappa$  and  $\mathbf{S}$  but also cope with ratio-valued data.

We briefly outline the general strategy behind the Bayesian simulation method. To keep the outline simple and within the scope of this article, we continue to treat the  $Y_i$  as real-valued random variables. A complete version is presented in Lee, *et al.* (2001). The population data  $\{Y_1, \dots, Y_N\}$  are taken as drawn from a parametric density  $f(y | \theta)$ , and  $y_1, \dots, y_k$  are a simple random sample from the population, and are treated as unobserved latent variables. Let  $\theta$  have a prior distribution  $\pi(\theta)$ . Denote the sum of the  $y$ -values in the  $i^{\text{th}}$  partition by  $y_{i+}$ :  $y_{i+} = \sum_{j \in \kappa_i} y_j$ . Then, based on aggregated data  $(n_1, y_{1+}), \dots, (n_k, y_{k+})$ , the likelihood is

$$\prod_{i=1}^k f^{*k}(y_{i+} | \theta) p(\kappa | \mathbf{S}),$$

where  $f^{*k}$  is the density of  $k$ -convolution of  $f$ . If  $f$  is a Gaussian or Gamma density,  $f^{*k}$  is known, but in many cases, of course,  $f^{*k}$  is unknown. This does not pose much difficulty in the Bayesian computation we propose.

The joint posterior distribution of  $(\theta, \mathbf{P} \setminus \mathbf{S}, \mathbf{S})$  given the aggregated data  $\mathbf{A}$  is proportional to

$$\pi(\theta) p(\kappa | \mathbf{S}) \prod_{i=1}^k f(Y_i | \theta) \prod_{i=1}^k I\left(\sum_{j \in \kappa_i} y_j = y_{i+}\right).$$

Note that the posterior distribution explicitly includes the term  $p(\kappa | \mathbf{S})$ , in order to model the data-dependent partition. Lee, *et al.* (2001) show how  $p(\kappa | \mathbf{S})$  could be specified..

This posterior distribution will almost always be intractable analytically. However, if we could generate a sample from it, we could: (1) Study the distribution of the population parameter  $\theta$ ; and then (2) Generate several realizations of  $\mathbf{P}$  and perform a bootstrap analysis (Efron & Tibshirani, 1986) on statistics of interest.

But, the posterior distribution will generally be too complex to allow direct generation of samples. We can, however, employ Markov chain Monte Carlo (MCMC) methods (Gilks, *et al.*, 1996) for the simulation. For MCMC, we simulate a particular Markov chain on the state space  $(\mathbf{P}, \theta)$  such that the stationary distribution of the chain is the posterior distribution.

Thus, by running the chain long enough, we can generate samples from the posterior. See Lee, *et al.* (2001) for details of the simulation procedure.

## 6. CONCLUSION

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 efficiently and effectively. Using it, NASS is able to provide information in more detail than previously, yet be assured that confidentiality criteria are fulfilled.

More complex data sets, queries and privacy concerns lead to additional challenges. For example, in many databases, the disclosure risk associated with a query depends on which queries have been answered previously. Integration of multiple databases raises additional issues, especially with risk computation and reduction. At the most basic level, a decision-theoretic formulation of the problem is necessary, in order to allow agencies to balance the value to society of releasing information derived from confidential data against disclosure risk.

## ACKNOWLEDGEMENTS

Carol House of NASS stimulated development of this system, and Joseph Prusacki of NASS provided both data and knowledge about them. Hassan Karimi of MCNC set up the GIS routines. We thank NISS summer interns Karen Brady and Christopher Holloman for numerous comments and suggestions. The research was supported by NSF grant EIA-9876619 to NISS. Jaeyong Lee is currently at the Pennsylvania State University.

### References

- Consumer Reports (2000). Who knows your medical secrets? *Consumer Reports*, **August** 23-26.
- Dalenius, T. (1977). Toward a methodology for statistical disclosure control. *Statistik Tidskrift* **5** 429-444.
- Duncan, G. T., de Wolf, V. A., Jabine, T. B., and Straf, M. L. (1993). Report of the panel on confidentiality and data access. *Journal of Official Statistics* **9** 271-274.
- Duncan, G. T., and Keller-McNulty, S. (2001). Mask or impute? *Review of Official Statistics* (to appear).
- Duncan, G. T., and Lambert, D. (1986). Disclosure-limited data dissemination (with discussion). *Journal of the American Statistical Association* **81** 10-28.

- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statistical Science* **1(3)** 54-77.
- Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology.
- Fienberg, S. E., and Willenborg, L. C. R. J., eds. (1998). Special Issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. *Journal of Official Statistics* **14(4)**.
- Gilks, W.R, Richardson, S., and Spiegelhalter, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.
- 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.
- Keller-McNulty, S., and Unger, E. A. (1993). Database systems: Inferential security. *Journal of Official Statistics* **9** 475-499.
- 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 survey. *Review of Official Statistics* (to appear).
- Schorr, H., and Stolfo, S. (1997). *Towards the Digital Government of the 21st Century*. Report from the Workshop on Research and Development Opportunities in the Federal Information Services. Available on-line at [www.isi.edu/nsf/prop.html](http://www.isi.edu/nsf/prop.html).
- Willenborg, L., and de Waal, T. (1987). *Statistical Disclosure Control in Practice*. Springer-Verlag, New York.

#### Appendix A: A Data Disclosure Primer

In broad terms, two kinds of disclosures are possible from a database of records containing attributes of individuals (e.g., Census records) or establishments (e.g., occupational safety data). An *identity disclosure* occurs when a record in the database can be associated with the individual or establishment that it describes. An *attribute disclosure* occurs if the value of a sensitive attribute, such as income or health status, is disclosed.

The first step in preventing identity disclosures is to remove explicit identifiers such as name and address or social security number, as well as implicit identifiers, such as "Occupation = Mayor of New York."

Often, however, this is not enough, because of the proliferation of databases and software to link records across databases. Record linkage produces identity disclosures by matching a record in the database to a record in another database containing (some of) the same attributes as well as identifiers. In one example, date of birth, zip code of residence and gender alone produced numerous identity disclosures from a medical records database by linkage to public voter registration data (Consumer Reports, 2000).

Identity disclosure can also occur by means of rare or extreme attributes. A data record for an eighty-year old Korean female dentist in North Dakota might easily be re-identified, as might Bill Gates' record in a database containing income of residents of Washington.

Aggregation (geographical or otherwise) is a principal strategy to reduce identity disclosures. The Census Bureau does not release data at aggregations less than 100,000. To prevent disclosing Bill Gates' identity by means of his income, all incomes exceeding \$10,000,000 could be lumped into a single category, a procedure called *top-coding*.

Attribute disclosure is often inferential in nature, and may not be entirely certain. For example, AIDS status (a most sensitive attribute) can be inferred with high certainty from

prescription records, but with less certainty from physician identity (if some physicians are known to specialize in treating AIDS).

Dominance can lead to attribute disclosure. The University of North Carolina at Chapel Hill is the dominant employer in Orange County, NC, so that the rate of workplace injuries for the county is, in effect, that for UNC. If this value is confidential at the establishment level, it cannot be disclosed at the county level.

There is a wealth of additional techniques (Fienberg & Willenborg, 1998) for “preventing” disclosure, which preserve low-dimensional statistical characteristics of the data, but distort disclosure-inducing high-dimensional characteristics. The  $(N, p)$ -rule employed in the NASS system is one example. *Cell suppression* is the outright refusal to release risky entries (typically, small ones) in tabular data. *Swapping* interchanges the values of one or more attributes, such as geography, between different data records. *Jittering* changes the values of sensitive attributes such as income by adding random noise. Even virtual databases can be created, which preserve some characteristics of the original data, but whose records simply do not correspond to real individuals or establishments (Duncan & Keller-McNulty, 2000).

#### Appendix B: FEDSTATS

More than 70 Federal agencies report expenditures of at least \$500,000 per year on statistical activities of collecting, analyzing and disseminating data. These include:

- **Bureau of Economic Analysis** ([www.bea.doc.gov](http://www.bea.doc.gov)): Statistics on gross domestic product, personal income and international trade;
- **Bureau of Labor Statistics** ([www.bls.gov](http://www.bls.gov)): Unemployment statistics, consumer price indices, occupational safety and health statistics;
- **Bureau of Justice Statistics** ([www.ojp.usdoj.gov/bjs](http://www.ojp.usdoj.gov/bjs)): Crime, victim, criminal offender and sentencing statistics;
- **Bureau of Transportation Statistics** ([www.bts.gov](http://www.bts.gov)): Highway safety, commodity and airline on-time statistics;
- **Census Bureau** ([www.census.gov](http://www.census.gov)): Population and economic statistics, especially from the decennial Census;
- **Energy Information Administration** ([www.eia.doe.gov](http://www.eia.doe.gov)): Statistics on energy consumption, cost and reserves, as well as projections of future usage;
- **National Agricultural Statistics Service** ([www.usda.gov/nass](http://www.usda.gov/nass)): Statistics on agricultural production and pesticide/herbicide/fungicide usage;
- **National Center for Education Statistics** ([nces.ed.gov](http://nces.ed.gov)): Statistics on educational achievement and education finance;
- **National Center for Health Statistics** ([www.cdc.gov/nchs](http://www.cdc.gov/nchs)): Statistics on births, marriages, divorces and deaths, prevalence of diseases, nursing homes and nutrition.

The Federal Interagency Council on Statistical Policy’s “FedStats” Web site ([www.fedstats.gov](http://www.fedstats.gov)) provides access to the full range of statistics and information produced for public use.

#### Appendix C: Algorithms for Aggregation

To describe the aggregation algorithms we use the following color code: **Red** = Undisclosable singleton county; **Pink** = Undisclosable supercounty (aggregate of singletons);

Blue = Undisclosable supercounty containing one or more disclosable singleton counties;  
 Green = Disclosable singleton county; Yellow = Disclosable supercounty.

The basic algorithm is:

**1. Start:** Color all disclosable counties Green and all undisclosable ones Red.

**2. Eliminate Red:** Examine each Red county according to a pre-selected random order. Merge with an appropriate neighboring county or supercounty, creating a Pink, Blue or Yellow supercounty. Continue until no Red counties remain.

**3. Eliminate Pink:** Examine each Pink supercounty according to a pre-selected random order. Merge with appropriate neighbors until no Pinks remain.

**4. Eliminate Blue:** Merge Blues with neighbors in the manner used for the Reds and the Pinks.

**5. End:** All (super)counties are Green or Yellow.

This procedure is guaranteed to terminate with a disclosable aggregation as long as the State-level data are disclosable.

What distinguishes the **small** and **pure** variants of the algorithm is how they define an appropriate neighbor to merge with. Table 1 lists the merging preferences for the two procedures, and should be read as follows. Consider, for example, the “Eliminate Red” section of the **small** algorithm. The preference is first to merge the Reds with other Reds to form disclosable (Yellow) supercounties. Whenever there are multiple candidates, one is selected either at random or on the basis of similarity of application rates. When all such cases have been exhausted, we next merge the Reds with any Blues that would yield a Yellow. We proceed down the list until all Reds have been merged.

It is clear from examining the table that **small** tends to produce small supercounties while **pure** favors leaving the disclosable counties alone and merging the undisclosable counties among themselves.

Table 1. Merging Rules for the **small** (top) and **pure** (bottom) algorithms.

Eliminate Red	Eliminate Pink	Eliminate Blue
R + R → Y	P + G → Y	B + B → Y
R + B → Y	P + B → Y	B + G → Y
R + G → Y	P + P → Y	B + Y → Y
R + P → Y	P + Y → Y	B + B → B
R + Y → Y	P + P → P	B + G → B
R + R → P	P + B → P	B + Y → B
R + P → P	P + G → P	
R + B → P	P + Y → P	
R + G → P		
R + R → P		

Eliminate Red	Eliminate Pink	Eliminate Blue
R + R → Y	P + B → Y	B + B → Y
R + B → Y	P + P → Y	B + Y → Y
R + P → Y	P + P → P	B + G → Y
R + R → P	P + Y → Y	B + B → B
R + P → P	P + G → Y	B + Y → B
R + Y → Y	P + B → B	B + G → B
R + G → Y	P + Y → P	
R + B → B	P + G → P	
R + Y → P		

Eliminate Red	Eliminate Pink	Eliminate Blue
$R + G \rightarrow P$		

The example in Figure 4 illustrates for a hypothetical small “State” made up of seven States in the Western USA. The data are displayed in the box beside each map. For the original data, WA, NV and NM are undisclosable and marked as Red. We first apply the **small** algorithm to produce an aggregation. Select a random order for merging the Reds, say {WA, NV, NM}. From Table 1, after the “Eliminate Red” step, we get the supercounties displayed in the second map. Note that the undisclosable NM has forced us to create a Blue region, wasting a disclosable county. We get the aggregation in the third map after the “Eliminate Pink” step and the aggregation in the fourth map after the “Eliminate Blue” step.

At this stage, all the regions are Yellow and hence disclosable. However, we see that {AZ, CA, NV, NM} can be decomposed into disclosable regions CA and {AZ, NV, NM}. This large region was created purely as a consequence of the visiting order we selected for the Red. Running the **pure** algorithm succeeds in breaking up the {AZ, CA, NV, NM} region, leading to the aggregation displayed in the final map.



Figure 4. Steps in the Heuristic Aggregation Procedure.