*Chapter Sixteen*

# Protection of Privacy through Microaggregation*

*Edgar L. Feige and Harold W. Watts*†

The enormous expansion of behavioral research coupled with what has been referred to as the information explosion has raised serious questions concerning the establishment, integration, and use of large-scale data collections, and the concomitant issue of the protection of privacy. The vast and rich data resources in the custody of governmental agencies, particularly at the federal level, present one of the most attractive fields of exploration for researchers. Without minimizing the problems associated with prospecting, assaying, and exploiting those data deposits, one must first find a permissible right-of-way through the no-trespassing signs which were erected to honor nondisclosure pledges made when the data

* The research underlying this paper was financed by a grant from the Board of Governors of the Federal Reserve System. The authors wish to express their thanks to Arthur S. Goldberger for his many helpful suggestions, comments, and criticisms.
† Social Systems Research Institute, University of Wisconsin, Madison.

were obtained and which must continue to be honored in order to defend the broad principle of protection of privacy.

We begin with the assumption that there does indeed exist a conflict between the principle of privacy and the need to discover new knowledge. Neither principle is regarded as absolute, and thus we can investigate the technical tradeoffs between the two. We specifically wish to investigate the costs associated with a particular scheme designed to protect privacy—namely, partial aggregation. Identities of individual reporting units and corresponding confidential information can be effectively obscured by combining the reports of several (or many) units into a single aggregated record. Partial aggregation replaces microunit observations, which are typically regarded as confidential, by nonconfidential observations on "statistical stereotypes." The observations on statistical stereotypes are constructed by combining microunits into groups and calculating the mean values for the grouped observations. If partial aggregation can be carried out without seriously impairing the usefulness of the data for legitimate research purposes, it can make a substantial contribution to our data base. Partial aggregation is not exactly a novelty; we have encountered it often in cross-tabulations, in parallel series on geographic subdivisions, and the like. Moreover, it has by these precedents been established as an acceptable means of preserving the confidential nature of the underlying data.

However, these familiar uses of partial aggregation may not adequately preserve the usefulness of the data. In particular they may be severely limited as primary inputs to regression analysis—probably the most widely used statistical procedure in social science studies. Although partial aggregation has been used in a variety of forms, there have been only a few studies that have explored the explicit costs of using partially aggregated data.* This paper is concerned first with establishing the properties of partially aggregated data as an input to regression analysis and secondly providing some preliminary results from an illustrative case study in which conjectures and procedures derived from an analytic investigation are given a trial on a major body of microdata.*

## PARTIAL AGGREGATION IN THE CONTEXT OF A SPECIFIC REGRESSION MODEL

In an earlier paper we considered the effects of partial aggregation on a specific regression model.[8] Instead of going into the technical details

* See Orcutt and Watts,[1] Bauman, David, Miller,[2] Miller,[3] Feige,[4] and Feige and Watts.[5]

* The analytical sections are indebted to previous work by Prais and Aitchison[6] and Cramer,[7] who have considered a similar problem from the point of view of reducing the magnitude of computations.

of that paper we shall simply outline the approach and summarize some of the conclusions reached.

After reviewing the classical multiple linear regression model and summarizing the conditions for obtaining best linear unbiased estimates of the regression coefficients, we proceed to show how disclosure can be avoided by the use of a $G$ grouping matrix that transforms the original microdata matrix into a matrix of group means. When the original data have been transformed by the use of the grouping matrix, the "generalized linear regression model" becomes appropriate, and the best linear unbiased estimator is the generalized least-squares estimator. Since it is typically very convenient from a computational point of view to utilize ordinary least-squares regression, we note that the generalized least-squares estimators can be obtained from an ordinary least-squares regression procedure when the microdata matrix is transformed by means of an $H$ grouping matrix that is related to the $G$ grouping matrix mentioned above. We then proceed to demonstrate that, given a grouping matrix, the least-squares estimator based on the partially aggregated data will be unbiased, providing that the grouping matrix is chosen independently and remains fixed over repeated sample realizations of the regression error term. This condition suggests that the choice of the grouping matrix is treated quite symmetrically with the choice of the independent variables in the regression model. It must be regarded as fixed in the same sense as the independent variables of the analysis. We can therefore ensure that estimators derived from partially aggregated data will be unbiased as long as the grouping matrix is chosen on the basis of the independent variables in the analysis, since the model assumes that the disturbance terms are identically distributed for all values of the independent variables.

If the grouping matrix is chosen on the basis of variables that have a systematic relationship to the disturbance terms, the resulting estimators will be biased. This means that if the grouping criterion is related to a variable that is omitted from the regression but systematically related to the disturbances, the estimator based on partially aggregated data will be biased.

The analysis goes on to show that partial aggregation does result in a loss of efficiency (increased sampling variability) for estimates of the regression coefficients and the residual error variance. The loss for the error variance is inevitable and proportional to the reduction in degrees of freedom. In the case of regression coefficients the loss depends on the nature of the grouping schemes. A trace-correlation index is defined that measures the relative efficiency of the grouped estimators as compared with the ungrouped estimators. In general it is shown that the loss of efficiency

from aggregation can be minimized by generating groups that are homogenous with respect to the independent variables of the analysis—that is, minimizing the within-group variation and maximizing the between-group variation. The relative efficiency index that we have constructed does not, however, provide an easy means of minimizing the loss of efficiency. The problem of classifying the microdata into groups so as to maximize the efficiency index is essentially an assignment problem and as such is not amenable to maximization via the familiar calculus methods. Nevertheless the efficiency index does provide a useful criterion for evaluating specific grouping schemes in the content of a specific regression model.

## PARTIAL AGGREGATION IN THE CONTEXT OF MULTIPURPOSE DATA

When one considers current data usage as well as the proposals for a federal data center one must focus particular attention on the use of partial aggregation of *multipurpose data*. It now becomes necessary to investigate the consequences of aggregating a body of data without reference to a particular regression model.

At the present time a large number of government agencies have collected highly useful data that have not been made generally available for research purposes. The major stumbling block to the wider dissemination of such data has been their confidential nature, insofar as they relate to individual microunits such as firms and individuals. In special instances agencies have used a variety of methods to obliterate identification characteristics of confidential data and have made some of the data available for special research projects; however, this procedure is both cumbersome and costly since each project requires special data tailoring. More generally government agencies have drastically aggregated microdata and have published these aggregated data for large geographic regions or for the country as a whole. The statistical limitations of such highly aggregated data have been widely discussed in the literature, and there appears to be a growing consensus concerning the greater usefulness of microdata or "slightly" aggregated data. In order to make existing confidential microdata accessible to the research community at substantially lower costs it is necessary to investigate the feasibility of designing multipurpose partial aggregation schemes that avoid the disclosure problem and yet have desirable statistical properties. Since the primary purpose of this chapter is to investigate the feasibility of constructing multipurpose aggregation schemes, it is necessary to

analyze the consequences of aggregating a body of data without reference to a particular regression model.

The aggregated data must avoid disclosure of confidential microdata and yet be amenable to the same range of uses as the unaggregated data. Ideally the aggregated data should have the following properties:

1. The relative loss of efficiency due to aggregation should be small for a wide range of regression models.
2. The estimators of the relevant population parameters applied to the grouped data should be unbiased.
3. The grouping scheme should be sufficiently flexible to permit the use of ancillary data that may or may not be subject to the disclosure problem.

The preceding analysis has considered the effects of data aggression on the estimators of a *specific linear regression* model. When the regression model is known in advance it is possible to devise grouping schemes that avoid disclosure of individual microdata and still maintain the property that the grouped estimators are unbiased. Moreover, given knowledge of the specific regression model and the microdata, it is possible to design a grouping scheme that results in relatively efficient grouped estimators. This latter objective can be accomplished by utilizing the independent variables of the analysis as grouping criteria and aggregating relatively homogeneous observations together in such a way as to minimize within-group variation and thereby maximizing between-group variation. The unbiased properties of the grouped estimators are ensured so long as the grouping scheme is chosen independently of the disturbance terms of the known regression model. For a specific regression model this condition rules out the selection of grouping schemes on the basis of the dependent variable of the model or any other variable that might be left out of the regression model, which is correlated with the disturbances. Operationally the above condition for unbiasedness is not difficult to fulfill. The variables to be used for the grouping criterion are the independent variables of the analysis, and, if the model is correctly specified, grouped estimators will be unbiased. However, when microdata are aggregated without reference to a *specific* regression model it is no longer possible to evaluate the effects of such aggregation on a priori grounds, nor is there any simple rule of thumb for "optimal aggregation." This is due to the fact that both efficiency and unbiased properties depend on the relationship between the grouping matrix and the disturbances in various regression models, which are not specified in advance.

If partial aggregation is to be used as a means of avoiding disclosure in the setting of the proposed federal statistical data center, then the problem raised by multipurpose aggregation are indeed significant. However, it may well be possible to design the operations of the federal statistical data center as to allow for *flexible aggregation* procedures. Such an approach would utilize a basic computer program that could be used for all analyses; however, the inputs to the computer would include prespecified preference priorities of the individual research worker requesting partially aggregated data.

An obvious case of prespecified preferences would arise where a researcher wished to avoid the assumption that behavioral parameters were the same for all groups of individuals in the population; for example, some investigators might wish to test the hypothesis that consumption coefficients differ between whites and nonwhites. If the data-organization scheme were perfectly rigid, partially aggregated groups would probably include both whites and nonwhites, and—although the average racial composition variable would vary from group to group, thus allowing some basis for discriminating different behavioral patterns—one might still prefer to treat the groups differently. Under a flexible aggregation scheme the computer could be instructed to segregate the population statistically into white and nonwhite groups and then to form partial clusters within each subgroup. Any prespecified preference that could be expressed in terms of a sorting operation could be included in this framework.

A similar problem arises when different researchers have widely different preferences concerning the relevant variables to be used for purposes of aggregation. It may well be possible to design a computer program that would utilize directly as inputs the specific preferences of the individual researcher. Included in the input would then be a listing of the variables that are to be used as the basis for clustering micro-observations.

As long as the variables to be used for clustering purposes are discrete in nature (i.e., income interval rather than income) the researcher can supply the computer with an *ordinal* set of grouping preferences that specify the priorities for consecutive sorts. The computer would of course reject any set of preferences that would result in less than a minimum number of individuals falling into any particular group.

More generally it is possible to inform the computer of the researcher's *cardinal* preference function with respect to the grouping criterion. The researcher in this case would have to regard the reduction of within-group variance of a particular variable as a scarce resource

and would have to indicate to the computer the price that he would be willing to pay—namely, an increase in the within-group variance with respect to some other variable. If the individual researcher would specify a tradeoff function, the computer could proceed to minimize within-group variances with respect to several variables, given the constraints imposed by the researcher's cardinal tradeoff function. Thus, if the researcher specifies that he is willing to increase the within-group variance of income by $100 in order to get a reduction of within-group variance of IQ level of 10 points, the computer would take account of this tradeoff and proceed to choose "optimal" groups in light of the prespecified quantitive preferences.

Such a flexible aggregation approach might possibly lead to some combination of sorting procedures that together could reveal some information about an individual. Although we suspect that this would be mathematically possible, it is not clear how such a scheme could be devised without some prior knowledge concerning individual units. If some information about a microunit is available through nonconfidential sources, it would be possible to gain added confidential information by use of successive sorting instructions. This possibility is much more likely to occur with firm data then with data on individuals. Since partial aggregation is not viewed as a unique panacea for all such problems, it may be quite possible to avoid such suggested violations of disclosure by raising the costs of such violations by other means, such as strict legislation.[9]

The more complex the data base becomes, the more important it may be to move in the direction of flexible aggregation schemes. Current procedures among government agencies typically involve rigid and drastic aggregation schemes. One might therefore wish to inquire how a rigid aggregation scheme would fare when the outputs of such a scheme become inputs to a variety of regression models. Moreover, we would hope to know more about the relative efficiency of estimators based on "slightly" aggregated data as opposed to "drastically" aggregated data. Finally, for a given degree of aggregation we would like to know the effects of different grouping criteria on the relative efficiency of resulting estimators.

In order to investigate these issues we undertook to examine a large body of micro analytical data that included call report and income-and-dividend information on over 5000 commercial banks that are members of the Federal Reserve System. The two groups of data came from separate sources and were linked by means of a bank identification code. These data were in turn linked with relevant ancillary data from the

city and county census tapes by means of geographic location codes, which were common to the two sets of data.

In order to evaluate the effects of aggregation schemes on a variety of regression models we formulated 20 experimental regression equations. Each one of these experimental equations was estimated by using the underlying microdata and then reestimated by using aggregated data that had been generated by a variety of aggregation schemes. For each regression model and for alternative aggregation schemes we computed an ex post facto index of relative efficiency that described the average loss of efficiency due to aggregation relative to the unaggregated ideal.

An aggregation scheme is defined both by the degree of aggregation (i.e., the size of the subgroups) and by an array rule that indicates which variables are to be used as the sorting criterion. One current Federal Reserve practice is to publish data on member banks that are aggregated to the state level. Such an aggregation procedure is drastic in the sense that it groups together as many as several hundred banks in some states and uses geographic location as the sole array rule. The index of efficiency for each of the experimental regression equations under this Federal Reserve aggregation scheme is displayed in Table 16.1. An efficiency index of unity indicates no loss of efficiency due to aggregation.

***Table 16.1***

***Relative Efficiency Index for State Aggregated Data***

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Regression number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Efficiency index | 0.199 | 0.081 | 0.160 | 0.153 | 0.220 | 0.211 | 0.195 | 0.112 | 0.230 |
| Regression number | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Efficiency index | 0.185 | 0.177 | 0.134 | 0.136 | 0.285 | 0.123 | 0.170 | 0.143 | 0.204 |
| Regression number | 19 | 20 | | | | | | | |
| Efficiency index | 0.038 | 0.037 | | | | | | | |

Average efficiency index over all regressions: 0.160

The results of Table 16.1 serve two purposes. First they indicate that the loss of efficiency from such drastic aggregation is substantial (its average efficiency index over all regressions is 0.160), and the loss of efficiency will, as expected, vary from equation to equation. In order to evaluate the effect of the degree of aggregation on loss of efficiency we used the same array rule (i.e., state and random within state) but computed the average efficiency index for different degrees of aggrega-

tion. In particular we considered groups consisting of 3, 30, and 100 banks. The resulting average efficiency indices were 0.446, 0.176, and 0.163, respectively. These results indicate that substantial gains in relative efficiency can be brought about by simply reducing the degree of aggregation.

Finally we experimented with using different array rules as criteria for aggregation. We found that some additional gains in efficiency could be brought about by using additional variables as grouping criteria. For example, when geographic location (state) was combined with a bank-size variable as an array rule the average relative efficiency index was raised to approximately 0.575 for "slight" degrees of aggregation.

The analysis strongly suggests that slight aggregation procedures, while still satisfying confidentiality requirements, can greatly improve our data base when compared to the current procedures of drastic aggregation. Looked at the other way, the results suggest that one cannot beat individual data when one's sole concern is with the highest statistical precision; but such precision might be well worth sacrificing to the principle of privacy. Moreover, aggregated data can be further improved by a judicious choice of array rules. On the basis of our analysis of bank data one is tempted to speculate that substantial gains in efficiency can be achieved by substantially reducing the degree of aggregation. Indeed, since the degree of aggregation appears to be more critical than the particular array rule chosen, one might argue that the gains to be achieved from flexible aggregation rules might not justify the marginal costs of such innovations. Although the rigid aggregation scheme does involve losses of efficiency in the multipurpose setting, it may be regarded as a useful starting point in the establishment of a federal statistical data center. A rigid aggregation scheme would not necessarily preclude using specially tailored aggregation schemes for particular high-priority projects that can demonstrate that the conventional mode of aggregation is simply inconsistent with the conceptual or statistical necessities of the project. Such special tailoring could be provided by means of the preference-function technique previously alluded to, as long as the data center maintained files on individual units.

The usefulness of a rigid aggregation scheme could be greatly enhanced if the data center also computed some of the common nonlinear transformations on the micro-observations before grouping those observations. In many research situations nonlinear models are specified but ultimately cast into the framework of linear regression models by utilizing transformations (squares, logs, reciprocals, products, etc.) on the

microdata. If the microdata have been linearly aggregated, transformations on the partially aggregated data can lead to serious problems of misspecification, due to the difference between the mean square (or product) and the square (or product) of the means or between the mean of the logs and the log of means. Heuristically it is clear that it is the within-group variation of the variables that enter into nonlinear transformations which causes difficulty. Linear aggregation as a means of avoiding disclosure can be utilized in the context of nonlinear models if all the transformations that may be needed are carried out before aggregation. Short of a complete solution to this problem, it might be useful to extend the basic set of variables to include commonly encountered transformations.

As a final speculative step we might ask what type of multipurpose aggregation scheme would be least offensive. We shall assume that actual data-linkage problems are solved by collating all data from different sources that relate to a particular microunit. Given a set of linked files of individual microunits we might select spatial location as the fundamental arary criterion. Geographic location has the advantage of being a universal characteristic since every microunit can be uniquely located spatially. It may well be possible to develop a numerical micro zip-code system which would have the property that aggregation of subdivisions would yield the numerical-code equivalent of larger subdivisions, and so on.

From the conceptual point of view geographic location may provide a useful aggregation criterion insofar as many behavioral models would wish to combine actors in a homogeneous environment, and spatial homogeneity is likely to serve this function quite well.

Finally, spatial location is likely to be associated with the independent variables in most behavioral studies and thus would satisfy one of the conditions required of aggregated data if biased estimators are to be avoided and efficiency losses kept to a minimum. The obvious exception to the foregoing would be migration studies, in which spatial location might be the dependent variable of the analysis and thus the worst criterion for aggregation. It would be possible to deal with such an exception by recourse to a specially tailored aggregation scheme of the type described above. In general we would expect that the geographic location criterion is best suited for the aggregation of firms, whereas a closely related scheme of constructing cohorts on the basis of birth year and sex may provide an excellent scheme for individuals. Under such an arrangement migration studies of individuals could use

cohorts as the inputs to the analysis since birth year and sex are not likely to show much variability.

## SUMMARY AND CONCLUSIONS

The great interest that has been aroused by the proposal of a federal statistical data center holds great promise and great problems. At the heart of the issue is the juxtaposition of conflicting values—the protection of privacy as opposed to the need to discover new knowledge. We have considered one possible means of partially reconciling these conflicting values—namely, the use of partial aggregation that obscures individual identification. Partial aggregation has the great advantage of satisfying all but the most severe of requirements for the protection of privacy and still offering great improvement over the data that are currently available to the research community.

The cost of adopting partial aggregation as a means of avoiding disclosure must clearly be measured in terms of the usefulness of partially aggregated data for research purposes. We have examined some of the consequences of using such data as an input to regression analysis and find that there are a variety of ways to reduce the potential loss of information due to aggregation. The most important gains may come from the adoption of "slight" as opposed to "drastic" aggregation schemes, and there do exist guidelines for the judicious choice of array rules that can further enhance the usefulness of partially aggregated data.

Partial aggregation cannot be regarded as a costless panacea for the many problems that have been raised concerning the proposed federal statistical data center, but it does provide a useful beginning toward reconciling some of the conflicting goals of such an establishment. It is surely possible to find some combination of statistical gimmicks and substantive legislative protections that can lead to both a maintenance of the privacy principle and an extension of our current data base and the concomitant expansion of our knowledge horizon. Although knowledge and privacy may be ultimately conflicting goals, it seems quite likely that, given the present state of both knowledge and privacy, a federal statistical data center, appropriately conceived and judiciously supervised, could indeed further both ends simultaneously.

REFERENCES

1. G. H. Orcutt and Harold Watts, *Consequences of Data Aggregation over Components for Prediction of the Effect of Policy on Economic Aggregates,* Systems Formulation, Methodology, and Policy Workshop Paper No. 6511. Madison, Wisc.: University of Wisconsin, September 1965.
2. R. A. Bauman, M. H. David, and R. F. Miller, *Sources of Income Variability for Wisconsin Male Taxpayers,* 1947–1959, Economic Behavior of Households Workshop Paper No. 6705. Madison, Wisc.: Social Systems Research Institute, University of Wisconsin, 1967.
3. R. F. Miller, *Confidentiality of Usability of Complex Data Bases,* Systems Formulation and Methodology Workshop Paper No. 6702. (Submitted to *The American Statistician,* May 1967.)
4. E. L. Feige, *The Organization of Financial Data for Research Purposes,* Financial and Fiscal Research Workshop Paper No. 6302. Madison, Wisc.: University of Wisconsin, October 1963.
5. E. L., Feige, and Harold W. Watts, *Partial Aggregation of Micro Economic Data,* Financial and Fiscal Research Workshop Paper No. 6601. Madison, Wisc.: University of Wisconsin, October 1963.
6. Prais and Aitchison, "The Grouping of Observations in Regression Analysis," *Review of the International Statistical Institute,* 1954, pp. 1–22.
7. J. S. Cramer, "Efficient Grouping, Regression, and Correlation in Engel Curve Analysis," *Journal of the American Statistical Association,* 59 (March 1964).
8. See ref. 5.
9. For an excellent discussion of these issues see ref. 3.