in Jerusalem, whose view is that only God can reestablish a Jewish state in Israel, and that a Jewish state established by human beings is a violation of God's will and so should be combatted. They see their mission as that of "guardians of the city," defending it from encroachment by secularity. As I read the evergrowing collection of papers authored or coauthored by David Freedman on the use of statistical procedures in modeling, I cannot help but dub him the "neturei karta," the "guardian of the city" of statistics.

How can one object to what he is trying to do? His quest, after all, seems quite reasonable. He tilts with models that are used in public policy deliberations and decisions. And he only concerns himself with the issue of whether the assumptions underlying the model are credible. Someone has to be the "guardian of the city!" Freedman is without peer in both thoroughness and clarity of analysis.

The problem, though, with Freedman's quest is in many ways analogous to that of the neturei karta. If they are successful, then the State of Israel will cease to exist. And if Freedman successfully uncovers models based on invalid assumptions, the decision maker is left to make decisions using only his intuition, for decisions must be made, with or without statistical help. All Freedman has done is saved statisticians from "aiding and abetting" and/or being accessories to a decision which in any event will be made, even if based merely on intuition and judgment. Is that worse or better than the scenario in which the statistician at least shows the decision maker the direction in which a decision should go, given the available data, in a (possibly) fictitious world built upon a bed of (possibly erroneous) assumptions? My contention is that even such deductions are useful grist for the decision maker's mill. Indeed, even if the assumptions are valid but the model is incomplete, or is just plain wrong, insights can be obtained from working the model through to its implied conclusions. (One can even gain insight from implications of purely mathematical models with no statistical component.)

Yes, assumptions should be checked for validity, and procedures should be checked for robustness. And no, statisticians are not merely people who "draw a straight line from an unwarranted assumption to a foregone conclusion using a procedure optimal according to a criterion invented by the statistician." But perhaps a bit of the latter can be condoned in statistical practice, especially if the alternative is that of letting the policy decision maker "go it alone." The statistician, after all, has more than a science to offer. He has a developed skill to offer as well, namely an ability to get the "feel" of data even when the data do not conform to any textbook model or set of assumptions.

ADDITIONAL REFERENCES

Cowan, C. D. and Bettin, P. J. (1982). *Estimates and Missing Data Problems in the Post Enumeration Program.* Statistical Methods Division, Bureau of the Census.

Freedman, D. (1981). Some pitfalls in large econometric models: a case study. *J. Business* **54** 479–500.

Freedman, D. (1985). Statistics and the scientific method. In *Cohort Analysis in Social Research* edited by W. M. Mason and S. E. Fienberg. New York, Springer-Verlag, pp. 343–366.

Kadane, J. B. (1984). Review of Mitroff, I. I., Mason, R. O. and Barabba, V. P. The 1980 census: policymaking amid turbulence. *J. Amer. Statist. Assoc.* **79** 467–469.

Mitroff, I. I., Mason, R. O. and Barabba, V. P. (1982). Policy as argument. *Manag. Sci.* **28** 1391–1404.

Mitroff, I. I., Mason, R. O. and Barabba, V. P. (1983). *The 1980 Census: Policymaking Amid Turbulence.* Lexington, MA, Lexington Books.

Ylvisaker, D. (1977). Test resistance. *J. Amer. Statist. Assoc.* **72** 551–556.

Comment

I. P. Fellegi

1. INTRODUCTION

I must state at the outset that I like the paper and would only have relatively unimportant technical "quibbles" to raise in *disagreement.* Instead, I will concentrate on some broader implications of the paper's findings. Another introductory comment is prompted by the paper's style, but applies to much of the written material on the topic of census adjustment. I would have preferred if the paper had more of a "sanitized" version of the authors' testimony, i.e., free of the debating style of courtrooms. The issues involved are both significant and complex, and it is all the more important that we should be able to debate our differences in a manner that makes it easier for our professional colleagues to understand our point of view, even if they disagree with it.

The paper clearly and, I believe conclusively, makes a case against a *specific* approach to adjustment. Yet its value goes well beyond its argument against a particular methodology. This is an important paper the careful reading of which imparts at the same time

I. P. Fellegi is Chief Statistician of Canada, Statistics Canada, R. H. Coats Building, 26-A, Ottawa, Ontario K1A 0T6, Canada.

both more and fewer lessons than its title implies. On the one hand, it doesn't really deal with "regression models" in their full possible generality. On the other hand, its careful analytical approach to model validation is a model (no pun intended!) for careful applied work. Trying to render explicit the hidden or implicit assumptions and then exploring, analytically and/or through data analysis and simulation, the impact of reasonable departures from them.

The controversy surrounding the issue of adjusting census data derives, in substance, from their basic importance. However, much of the *statistical* controversy can be traced to the fact that statisticians are trained to want to exploit all available information. We know from several census evaluations that important differential undercounts exist, for example, by age, sex, race, and geography. It goes against the grain not to use this information. Yet, under certain conditions formal utilization of available data can be harmful. The paper makes the important point, if indirectly, that the range of uncertainty (both variance and unknown bias) of Post Enumeration Program (PEP) estimates is sufficiently large *relative to likely errors in the census* that its use in 1980 to adjust the census data might well have been harmful. This general point can be illustrated in a trivial (and obviously unrealistic) example. Let x and y be independent random variables with the same expected value. Their variances are unknown but we know that x has substantially smaller variance. Now to estimate $E(x) = E(y)$ it is quite clear that there exists linear combinations $ax + by$ $(a + b = 1)$ the variance of which is larger than that of x. Under such conditions, there may well be circumstances under which ignoring y may not be a bad strategy. The example would have to be *substantially* elaborated to be illustrative of the complexities involved in adjusting census counts (e.g., x and y are both biased and $E(x) \neq E(y)$), but it might serve to illuminate the frustration in having data which one may not be able to exploit formally because of its unknown error properties.

The paper has important implications with respect to a program that might lead to a capability in the Bureau of the Census to adjust the "raw" census counts. I shall leave aside issues of logistics (the adjustment, to be useful for congressional seat allocation, would have to be accomplished before the Census Bureau's legally mandated deadline of December 31, 1990), law, and public perceptions. The following three broad technical prerequisites would have to be met: a substantially improved coverage evaluation program would have to be developed; models would have to be identified capable of producing "good" coverage adjustment factors for some limited number (100?) of large areas; and "good" methods would have to be developed to carry down the large area adjustments to smaller areas. A few comments, in the nature of personal speculations prompted by the paper, are in order about each of these issues.

2. COVERAGE EVALUATION ISSUES

The paper clearly illustrates the central role of the coverage evaluation program. Three issues emerge.

(a) The validity of the assumptions underlying any subsequent modeling work clearly depends on the error structure of the input data, in the present case the coverage evaluation results. Section 4 presents the particular set of assumptions required for the hierarchical Bayes method, as put forward by New York, and Section 5 provides a critique of the applicability of these assumptions in light of the 1980 PEP program. Different models may, of course, be developed before 1990, involving different assumptions. However, it is almost certain that strong assumptions will be needed if reasonably complex models are to be tractable. But the error structure of population survey data, such as PEP, are notoriously difficult to parameterize. I believe that the likeliest way out of this conundrum is *not* to try to capture the exceptionally complex error structure of a survey like PEP through increasingly complex models. Rather, we have to seek models which are exceptionally robust with respect to deviations from the assumed error behavior of the input data. This may, of course, mean sacrificing elegance or optimality.

(b) The second issue relates to basic improvements needed in the PEP program to be able to support a subsequent adjustment. This is probably the single most difficult prerequisite—the Achilles heel of adjustment strategies. Let us remind ourselves of an important fact. PEP is designed to estimate the bias (due to coverage errors) of census counts, and that this bias is reasonably small (say, less than 2% overall, up to about 5% for blacks). So when we are talking about errors in PEP, we are dealing with the variances and biases of an estimated small bias! Table 2 of the paper shows the high level of indeterminacy due to differential biases in two sets of PEP data. Both sets are plausible, and the differences arise as a result of the treatment of such factors as nonresponse, matching errors, and timing of the reinterview date. There are certainly possibilities for improving the PEP data in 1990 through improved survey design, more follow-up of nonresponse, automation of matching routines, and the like. Yet, the source of the needed *breakthroughs* is not visible yet.

(c) A major new problem of the "evaluability" of coverage of U. S censuses is due to the high level of undocumented aliens which is of the same order of magnitude as the coverage error itself. The intended coverage of the 1980 Census was to include them.

There is likely to be a very much higher level of census undercoverage in this group compared to the legally resident population. The same applies to the reinterview surveys. In fact, it is probable that a significant but unknown proportion of them has a zero probability of inclusion in either the census or PEP. But a basic assumption of capture-recapture methods, such as PEP, is that each member of the population should have a nonzero probability of inclusion in both survey and census. Large biases can arise if this probability is zero for a significant proportion of the population (significant relative to the coverage error which is to be estimated). In some respects it is even worse if those having close to zero probability of inclusion in the census have positive inclusion probability in PEP or vice versa. All of these contingencies are not only possible but likely.

A *partial* solution of this problem could be developed if the census and PEP could identify separately the legally resident population. However, the census question needed to achieve this objective would probably be regarded as an unacceptable invasion of privacy.

3. ISSUES RELATED TO MODELS

The paper highlights some very difficult model development issues. Mention has already been made of the difficulty of making acceptable assumptions with respect to the error structure of PEP data. Additional difficulties arise because PEP data cannot be assumed to provide unbiased estimates of coverage errors as illustrated by equation (23) of the paper. Models would have to be developed that make allowances for these biases. Perhaps their impact could be made explicit within the model algebraically and "conservative" adjustment strategies developed by making reasonable assumptions about their direction, somewhat in analogy to the work of Fellegi (1981).

Another problem highlighted by Tables 3 and 5 of the paper relates to the choice of independent variables. Clearly, no small subset of the nine independent variables could be identified which would adequately model the dependent variable. Collinearity does not appear to be the main reason, but rather the inadequate explanatory power of any of the sets of three variables considered. Yet the development of an authoritative set of independent variables is important since someone could make a significant improvement to the model by including additional variables no matter what variables were chosen by the Census Bureau. This reflects, of course, the substantive complexity of the phenomenon of undercoverage and the consequent difficulty of modeling it.

Notwithstanding the modeling difficulties, intensive research is, I believe, more likely to be productive here than efforts to overcome problems having to do with the quality of PEP or other coverage evaluation data.

4. ADJUSTING FOR SMALL AREAS

There are a multitude of needs for small area data; general revenue sharing alone requires population counts (*and* income data) for some 39,000 areas. Yet, given the relatively low level of undercoverage, even a large evaluation survey like the 1980 PEP can only provide direct estimates with acceptable sampling error for a relatively small number of areas, almost certainly less than 1,000 and probably more like 100. New York's methodology, which is criticized by the paper, uses modeling to try to reduce the error of the direct PEP estimates for these areas. Alternative strategies based on a different design for PEP than that used in 1980 could generate model-based estimates for a somewhat larger number of areas without extrapolating outside the model. Table 4 of the paper and the discussion surrounding equations (11)–(14) certainly illustrate the extreme risk involved in such extrapolation.

Assuming that a good adjustment model could be developed for larger areas and that the problems of the quality of coverage evaluation data could be solved, an alternative strategy would be to adjust the component small areas using a methodology which is less likely to generate outliers than a direct extrapolation of the model. Synthetic estimates or iterative proportional fitting are in this category and other strategies could be devised designed to attenuate or censor particularly large adjustments of the original census counts.

Whatever the proposed methods for "carrying down" the adjustment, the paper illustrates that it is exceedingly important to test them in light of empirical information. This has major implications for the design of a PEP program. It should facilitate the validation of proposed methodologies by providing, for a sample of "small" areas, direct estimates of the undercount.

5. CONCLUSION

This important paper is a major contribution to the public debate regarding the feasibility of adjusting census data. By highlighting pitfalls, it points the way to future research which, if successful, might lead to effective adjustment methodologies, including developments needed in both models and input data. The Bureau of the Census should be congratulated for having made its PEP data available in a mode to facilitate this type of secondary analysis, for example, by making explicit a range of potential and feasible imputations in PEP. Last but not least, Ericksen and Kadane have shown courage and innovation by putting forward a methodology in an area fraught with extreme difficulty.

For a broad range of uses the census data are accurate enough, like Newton's laws prior to the discovery of the theory of relativity. A higher intended standard of accuracy, deriving from one man one vote principles and large fund allocations tied to census results, seem to demand a new level of precision. Yet, we have not evolved the needed "theory of relativity" in the area of census adjustment, nor the statistical measuring instruments which could serve as yardsticks when approaching the speed of light. Parenthetically, given the very high level of intercensal mobility and the relatively crude methodology available to track it, it is not entirely obvious why the census must have such extraordinary point-in-time precision. Indeed, over a decade the most disadvantaged areas in terms of congressional representation are undoubtedly those having the highest growth rate.

I am not optimistic about the likelihood of overcoming the technical difficulties involved by 1990, but the issue is clearly important enough so that a major effort must be made.

ADDITIONAL REFERENCE

FELLEGI, I. P. (1981). Should the census be adjusted for allocation purposes? Equity considerations. *Curr. Top. Survey Sampling* 47–76.

Comment

Lincoln E. Moses

This paper shows one side of an argument between two sets of statisticians. The argument was a court case between the country's biggest city and the federal government, with many millions of dollars at stake. No wonder it is fascinating reading. Perhaps it is more surprising that upon reflection I find this paper very convincing, even though I have read just this one side.

Convincing and important.

Freedman and Navidi first describe the census, the Post Enumeration Program (PEP) series, and the approach of New York City to estimating census undercounts by regression of PEP estimates on a number of demographic covariates for 66 areas.

Then they lay bare the assumptions on which depends the validity of the analysis offered by New York City. There are seven such assumptions and the authors give us ample reason to doubt each one. Theorems, real-world heuristics, computations, and experimental sampling are all drawn upon, leaving this reader persuaded that New York City had little claim to having shown a way to improve the census figures by means of regression adjustment.

Freedman and Navidi show that some assumptions are implausible on their face (for example, the independence of two kinds of error component, and that variance of one of them could be regarded as known.) They establish that the model entails the assumption that bias in the PEP figures is *not* related to the very demographic variables that are supposed to account for much of the bias in the census, the variables that are to be used to correct the census bias (undercount). They comment on the implausibility of this assumption, and then construct a second series of PEP adjustments, rather parallel to the series used by New York City and find that the difference between the two adjusted series is highly correlated with the demographic variables, which implies that at least one of the two PEP series must fail the key assumption that bias in PEP be unrelated to the demographic variables. The argument to this point implies that biases (assumed away by New York City) are likely operating, making standard errors inadequate measures of error. Then, by means of bootstrap sampling emerges the empirical information that indeed the New York City standard errors (given by formulas appropriate to the theoretical model) do understate the mean square error obtained by empirical sampling from a model in which many of the assumptions by New York City were made true by construction.

Freedman and Navidi have not attacked a strawman, they have not simply set out to find flaws in an example, they have assumed the burden of showing that New York City has not shown how to use the PEP estimates, plus regression, to give improved census counts. If they have succeeded in this (as I think), why is it important to statisticians?

First, statistical argument is becoming more frequent in litigation, so our profession is learning by doing. This case is an instructive example; it shows

Lincoln E. Moses is Professor of Statistics in the Departments of Statistics and Family Community and Preventive Medicine of Stanford University. His mailing address is Department of Statistics, Stanford University, Stanford, CA 94305.