Journal of Official Statistics
Vol. 2, No. 4, 1986, pp. 531–544
© Statistics Sweden

Census Data Quality – A User's View[1]

William J. Hawkes, Jr.[2]

Abstract: This paper presents the perspective of a major user of both decennial and economic census data. It illustrates how these data are used as a framework for commercial marketing research surveys that measure television audiences and sales of consumer goods through retail stores, drawing on Nielsen's own experience in data collection and evaluation. It reviews Nielsen's analyses of census data quality based, in part, on actual field evaluation of census results. Finally, it suggests ways that data quality might be evaluated and improved to enhance the usefulness of these census programs.

Key words: Evaluation; census quality; mean-square-error.

1. Introduction

The A.C. Nielsen Company is a major user of periodic census data – both the decennial Census of Population and the quinquennial Census of Retail Trade. In our dual role of statistics producer as well as statistics user, we have had the opportunity to do our own evaluation of census data quality, to experience and understand the effects of census quality on our own reports, and to apply what we have learned about data quality to our own operations. We have done this from the unique perspective afforded by evaluating census data quality in each of the twenty-five countries in which Nielsen conducts its marketing research activities. From this vantage point, we have come to appreciate the value of dependable, high quality census data that are the result of the U.S. Census Bureau's institutional history, its tradition of adherence to the highest professional standards, and its covenant of trust with the American people, reflecting both the confidentiality requirements of U.S. law and the dedication of the Census Bureau in supporting and upholding these requirements.

One of the primary marketing research uses of census data – both economic and demographic – is to assist in establishing a sampling frame and a projection base for marketing research surveys. This paper describes briefly how census data are utilized by Nielsen for marketing purposes.

2. Use of Decennial Population Census Data for Nielsen Television Ratings

Nielsen national television ratings in the U.S. are based on continuous electronic measurements of television set-tuning in a sample of around 1 700 households across the United States. In an additional nine individual major markets (Nielsen Metered Television Markets (1985)), similarily metered measurements are carried out in samples of around 500 households in each market.

[1] An earlier version of the paper was published in the Proceedings of the U.S. Bureau of the Census First Annual Research Conference, March 20–23, 1985.

[2] Vice President, Director of Statistical Research, A.C. Nielsen Company, Northbrook, Illinois, USA.

With the release of each new decennial census, a new Nielsen master sample is designed and selected. This sample is kept up-to-date between census years by surveys based on Census Bureau information on new building permits, and by periodic resurveys of the master sample frame in areas of the country not covered by building permits (approximately 12 % of the U.S.).

The Nielsen television master sample is selected by means of: a four-stage area probability sample design using decennial census data, very small geographic areas, such as blocks or enumeration districts, and Nielsen sample specifications. These geographic areas are then canvassed in their entirety by Nielsen interviewers. This survey yields sample household names and addresses on a probability basis without requiring that these household names and addresses in any way be supplied from census records.

The first stage of sampling involves what are called primary areas. A primary area consists of one or more contiguous counties and contains a minimum of 5 000 housing units, generally based on standard metropolitan statistical area definitions and on individual counties or contiguous county combinations elsewhere. Contiguous primary areas are then grouped into clusters containing around 225 000 housing units, with one or two primary areas sampled from each cluster using a "controlled selection" procedure to assure sample representation by stratification criteria such as county size and percent of households subscribing to cable television. Primary areas over approximately 180 000 housing units are selected with certainty.

The second stage of sampling involves the selection of "block groups" or enumeration districts. This is also accomplished using a controlled selection procedure using the marginal constraints or stratification variables of Nielsen territory, county size, number of households that subscribe to cable TV, percent minority households according to the 1980 Census, and percent households with children, again according to the 1980 Census. Within each of these strata, a random or systematic selection of sample block groups or enumeration districts are made.

The third stage of sampling involves selecting a sample of block segments from a listing of all such segments arranged in geographical sequence within the sampling units chosen in the second stage of sampling.

The small geographic segments selected in this way are then canvassed in their entirety by Nielsen surveyors. In the fourth stage of sampling, individual sample households are chosen from the housing unit frame delineated by this enumeration.

The sample design described above is reasonably robust against errors in the basic census materials from which the first three stages of the sample are selected. Any inaccuracy in housing unit census counts for small areas will obviously increase the variance of our sample-based measurements. Moreover, certain demographic characteristics derived from the Current Population Survey are used as universe controls to which individual household data will subsequently be projected. This means that it is important that population and housing unit counts, as well as population characteristics, be obtained and reported by the Census Bureau using definitions and procedures that are consistent and reproducible. A few problem areas that we have encountered in this regard will be discussed later in this paper. We first turn, however, to our use of economic census data.

3. Use of Quinquennial Economic Census Data for Nielsen Food and Drug Indexes

For the past 50 years, the A.C. Nielsen Company has provided the leading manufacturers of consumer goods with factual information on the sales of these products in retail stores, primarily through grocery and drug stores. This information is based on actual field audits conducted in national samples of around 1 300 grocery stores and around 700 drug and proprietary stores. These sales data are also supplemented by additional samples in individual metropolitan or local geographic areas that manufacturers utilize for new product or advertising testing purposes. The design of these samples is partly based on information concerning the retail grocery and drug store universe generated by the Census Bureau through its quinquennial Census of Business, supplemented by more up-to-date information contained in County Business Patterns and the Current Monthly and Annual Retail Trade Surveys.

Nielsen Food Index and Drug Index samples have been selected using a two-stage sampling process. The first stage designated the counties in which sample stores are located and the second stage designated the sample of stores within these counties.

In the first stage of the selection process, all counties included in the 38 Nielsen major market areas and all additional metropolitan areas with a 1980 population of at least 350 000 were designated as "certainty" counties. These 1 133 designated counties include about 73 % of the U.S. population. The remaining 1 940 counties were classified into 198 sample clusters. Clusters are formed by geographically contiguous counties, with each cluster including an average of ten counties and a population of 309 000. Within each sample cluster, a sample county was selected with a probability proportional to the 1980 Census population counts.

In the second stage, each grocery and drug store universe was divided into approximately 150 mutually exclusive strata, with each stratum defined by geographic region, county size, store type and size characteristics. Information on the number of stores in each stratum and the all-commodity sales importance of these stores was obtained from the most recent U.S. Census of Retail Trade (through special tabulations prepared for the A.C. Nielsen Company). Sample sizes by store type and size were based on an optimal allocation procedure. Within each stratum, individual sample stores were selected using a geographically sequenced sampling procedure that selected individual counties based on information from the U.S. Census of Business. For the chain strata, additional controls were imposed for each major chain organization. Within the food chain and large independent strata, and for all drug stores, sample stores were selected from a current universe listing drawn from commercial sources, individual firm lists, and Nielsen surveys. Smaller independent food stores were specified by county and store size, using U.S. census data, although name and address locations must be obtained in the field.

For observations at the individual store level, there is generally a high correlation between sales of an individual grocery item and the store's total sales. We make use of a ratio estimation projection procedure in which, for each stratum, universe all-commodity dollar sales as derived from U.S. census data are divided by the sum of all-commodity dollar sales for all establishments included in the Nielsen sample. In this way we establish a ratio estimator by which individual item sales in sample stores are multiplied for reporting purposes.

Because of changes constantly taking place within the retail universes, it is necessary to update continously the sources of information on the universes. This is accomplished using: the Current Population Survey, Annual Retail Trade Survey, and the County Business Patterns Program, as well as Nielsen's own surveys. Both the County Business Patterns and the Current Retail Trade Survey are based, in part, on the updated quinquennial census frame, and the standard statistical establishment list. Thus, certain errors in concept, definition, scope, or classification in the quinquennial census data will be carried over to the more current estimates and will introduce errors into Nielsen data. Here, too, as with the population census, we need to be acquainted with all census procedures and potential error sources so that whenever possible we can modify our sample and projection base accordingly.

4. Comments on Evaluating Census Data Quality

We will now consider certain quality aspects affecting our use of census data. It is important to recognize that the quality of census data, especially the population census, has been the subject of intensive evaluation, both within and outside the Census Bureau, for many years. Probably no statistical undertaking has ever been subject to such close scrutiny, nor the topic of so many research studies, as the last several decennial censuses. It is important, also, to credit the Census Bureau itself for having pioneered the tradition of such evaluation and review, to the point that at times the bureau has seemed to be its own severest critic – and indeed to have invited external criticism through publication of its own quality evaluations.

Such high standards of professionalism are admirable, although difficult to emulate. But they also make it more difficult for an "outside" user to make a significant contribution to the evaluation process. There is relatively little new knowledge that one can add to the content evaluations, response bias and response variance appraisals, or procedural and analytical disciplines that have already been carried out and described by the Census Bureau.

It does seem important to stress that quality control procedures and evaluations must be systematic if they are to be effective. One must recognize the interrelatedness of individual quality aspects, and then monitor all links in the quality chain. This approach, long an objective of the bureau's demographic census evaluation, is now being extended to economic areas as well.

Accordingly, I will not attempt to provide a comprehensive evaluation of census data quality, or even a comprehensive taxonomy of evaluative criteria or sources of error. Nor do I wish merely to discuss the bureau's quality evaluation programs as the bureau has described them. Instead, I will focus on a few quality aspects that we have experienced as users of the data and on a few quality aspects that we have encountered in certain enumerative efforts of our own. Finally, I wish to propose consideration and evaluation of an error component that is crucial for the user of census data to understand, but which has received little attention – and even less attempt at measurement – in the literature.

5. Evaluation of Decennial Population Census Results

As described in Section 2 above, Nielsen field enumerators, in the course of compiling the master sample frame for our television index sample, canvass a large number of geographic land area segments following each decennial population census. Around 8 400 of these segments (generally contiguous blocks or block groups) makes a total of around 2 180 000 housing units that were canvassed during the summer of 1982. We recognize that this was around 27 months after the reference date of the 1980 Census and that certain changes in the housing stock occurred during this period. In our analysis we have taken account of the construction or demolition that could be identified or documented. Many segments had changed very little, if at all, in the interim. We had the opportunity to conduct our own "post-enumeration surveys" or, if you will, our own "mini-census" for these areas. While this survey or census enumerates only housing units, it does provide an independent check on the completeness of census housing units.

What we learned from this was of value to both the Census Bureau and the Nielsen Company. We learned that it is easier to take a census if you are the Census Bureau than if you are the Nielsen Company. We were denied physical access to some of the housing units because they were in multi-unit structures that we were unable to canvass. At other units there simply was no one at home. However, for the most part, we were able to gain information on housing units from someone other than a household member, like a neighbor – but some housing unit determinations were, of necessity, conjectural.

Even after allowing for nonresponse bias we could not find as high a proportion of one-person households as was shown by the Census Bureau. Part of this difference may reflect the ability of post office carriers, in the census precanvass, to know how many multiple households – or, more accurately, how many housing unit "repartitions," there actually are at an address. To the observer (or to the zoning board) these multiple households can often look like a single housing unit. This would seem consistent with the Census Bureau's own conclusion from its 1970 CPS-Census match that there was "better coverage of housing units in mail areas than in conventional (i.e., listing) areas." In mail areas, lists compiled from conventional registers appeared to produce better coverage than prelisting did. (U.S. Bureau of the Census (1978, p. 39).)

Our experience in identifying housing units makes us keenly aware of how tenuous the notion of "usual place of residence" actually is – especially in the fluid, mobile society of the 1980s. For example, how is the manager of an apartment complex that rents by the month to know which of his tenants consider that apartment to be their "usual place of residence"? What about weekend or vacation homes? Or two wage-earner married couples with distant jobs and (sometimes) separate living arrangements? Or young executives on a six-month management training assignment in another city? One might find considerable response variance, from day to day, with the same respondent as to whether he/she considers himself/herself a separate household at the time of the inquiry. This is not to suggest that we abandon the "housing unit" concept, but simply that we recognize its complexity and even its stochastic nature. As noted, not all constructs are inherently coherent or even measurable.

Our studies also support the conclusion reported by the Census Bureau that population census results reflect considerable geographic miscoding by block (8.4 % of all units in blocked areas were reported as miscoded in the 1970 Census) (U.S. Bureau of the Census (1978, p. 55)). While most of these were coded in the right tract, this phenomenon sometimes causes gross understatement or overstatement. Since the household's final probability of selection in the NTI sample is based on our own field canvass results and not the census count, these errors do not bias our sampling frame. These errors are, however, a source of additional variance.

We were surprised to find that the quality of census maps used in 1980 was often inferior to the maps used for the 1970 Census. While the enumeration district maps in rural areas were good, the block assignments in the metropolitan map series left much to be desired. Some blocks were left un-numbered, while other block numbers were mysteriously repeated. This undoubtedly contributes to the geographic block miscoding reported by the 1980 Census. Mapping in metropolitan areas, at least through 1980, has been carried out by clerical means. The process has been strongly resistant to technological improvement or automation.

On the whole the census housing unit counts and ours are reassuringly close, as shown in Table 1.

Table 1. Comparison of A.C. Nielsen 1982 Field Canvass of Housing Units With 1980 Census Housing Unit Counts by Block Group or Enumeration District (National Nielsen Television Index Survey Segments Only)

Differences from 1980 Census Count (percent)	Number of Survey Segments
+ 30 or more	74
+ 20 to + 30	40
+ 15 to + 20	33
+ 10 to + 15	62
+ 7.5 to + 10	50
+ 5.5 to + 7.5	71
+ 3.5 to + 5.5	90
+ 2.5 to + 3.5	67
+ 1.5 to + 2.5	76
+ 0.5 to + 1.5	119
− 0.5 to + 0.5	181
− 1.5 to − 0.5	156
− 2.5 to − 1.5	115
− 3.5 to − 2.5	105
− 5.5 to − 3.5	153
− 7.5 to − 5.5	145
− 10.0 to − 7.5	112
− 15.0 to − 10.0	133
− 20.0 to − 15.0	65
− 30.0 to − 20.0	86
Less than − 30.0	68
All segments	2 001
Median difference	− 0.4%
Median difference, without regard to sign	± 4.9%
Average number of housing units per segment	270

We did experience some difficulty, as did many census users, with the large increase in the proportion of persons classified as "other races" in the 1980 Census. Our conjecture is that the exclusion of the words "race" or "color" from the census question caused confusion. The 1970 Census did include them, and without these words in the question, a respondent might well wonder what the appropriate response context was. While it is true that more persons of Spanish origin classified themselves as "other" in 1980 than in 1970, the questionnaire itself may have contributed to this reclassification.

6. Evaluation of Economic Census Data Quality

The accuracy of consumer sales information provided by Nielsen Retail Index measurements is often appraised independently by our manufacturer clients and by examination of sampling variability. Since most of the merchandise produced by a manufacturer will eventually be bought by a consumer, it is a relatively straightforward process for a manufacturer to compare his factory shipments over a period of several years with the consumer sales reported by Nielsen. Our ability to track consumer sales movement within specified tolerances will be validated in a practical way. We need to be sure that the census-based all-commodity dollar sales figures that our sample store activity is projected to by stratum (as described earlier) are complete, up-to-date, and measured by the census in a way consistent with the way we measure them.

Some factors influencing the quality and comparability of economic census data are outlined in subsection 6.1 through 6.6.

6.1. Clarity and reproducibility of the concept of "establishment"

Just as there is some ambiguity inherent in the concept of housing unit or household, the concept of "establishment" is not always clear to the respondent. For example, is a supermarket – drug store combination, operated by two sub-groups of the same retail firm, with common entrances, common checkout facilities, but with both names on the storefront and separate bookkeeping, considered to be one establishment or two? Who decides? How much information does that decision maker have regarding the Standard Industrial Classification (SIC) "establishment" concept? If the decision is made by the organization to fill out separate forms and thereby report two establishments, is that decision made with full knowledge of census intent on how such establishments "ought" to be classified? Does the census have enough information on the firm's activities to overrule the firm's decision? Are these decisions consistent through firms and across time within the individual firm? Are they consistent between the census and Nielsen?

Similarly, there tends to be an "agglomeration" bias on the part of smaller multi-unit firms in providing the census with a single combined report for what is actually two or more establishments. This can create incorrect measures of size and incorrect strata allocations by Nielsen. It also makes it impossible to "replicate" census counts of the number of establishments in a particular geographic area, resulting in unexplained discrepancies between the census-defined universe and the Nielsen-enumerated establishment sampling frame designed to provide access to that universe.

6.2. Definition and reporting of sales receipts

As Oscar Morgenstern (1963) and others have pointed out, economic data at their source, i.e., the individual firm or establishment, sometimes lack the exactness implied by the number of digits with which they are reported. "Sales," "receipts," and "payroll" seem simple enough concepts, but both conceptual and measurement errors can occur in their compilation, estimation, or reporting.

In general, we have found that U.S. retailers keep good sales records and report them accurately and honestly. Most confusion or uncertainty concerns a limited number of items and a relatively few retailers.

One exception to this generalization is the inclusion or exclusion of state and local sales taxes. This has been identified by the Census Bureau as a source of fairly widespread error (or inconsistency), although it is one of the few sources of error addressable through clearer instructions (U.S. Bureau of the Census (1984)). Our experience with retailer records certainly confirms this finding. It is made particularly difficult because individual states vary in their tax reporting requirements, procedures, and in the way that taxes are recorded at the cash register.

Handling of credits, returns, and carrying charges is another frequent source of error. Other sources of error are the numerous components of income generated by non-retail sources in retail firms and income generated by non-service sources in service firms. Often the need for quick reporting works against the need for accurate reporting, and even "book" figures are at times provisional.

6.3. Geographic coding of establishments

We have found that economic census data are subject to a "name of place" bias because many businesses are located at sites different from their mailing addresses. Many retailers tend to think of their physical location in terms of their postal addresses, even though they actually are located somewhere else. This phenomenon does not affect county or state-level data, but should be kept in mind when working with "place-level" aggregates.

It is difficult to explain to the occasional user of census data why 80 million households can be assigned to census tracts and even city blocks while seven million business establishments apparently cannot. Yet anyone who tries to match address-coded lists of businesses – or locate them in the field – soon recognizes the difficulty. "Prestige addresses," shopping center locations, highway designations, and the place-name bias all complicate the process. Even the Nielsen Company is a part of the problem, since it is probably impossible to find "Nielsen Plaza, Northbrook, Illinois" on any map used by the Census Bureau.

6.4. Limitations of the standard industrial classification system

Because the economic census is based on self-enumeration by mail, and because the standard industrial classification system cannot anticipate the emergence of new types of business firms, census data have been largely silent on certain kinds of businesses, such as mass merchandisers[3] and computer stores, that are of considerable interest to the marketing community.

As an example, the familiar retail phenomenon generally known as the "mass merchandiser" has yet to find a place in the standard industrial directory. Only through innovation on the part of the Census Bureau has it been possible for this kind of business to be reported, for the first time, in the recently published 1982 Economic Census. Consequently, each year for the past twenty-three years Nielsen has had to conduct its own annual "census" of mass merchandisers in order to provide a basis for sales measurements within this type of outlet. Starting this year, we are inaugurating a new service measuring sales in computer stores, which required that we conduct our own census on this new retail category as well. Computer stores are not yet reported in any census publication, and are not likely to be reported for some time.

Another limitation of the standard industrial classification system is that certain retail activities conducted by service firms, such as eating and drinking places operated by hotels and motels, are counted with Service Trades, but not with their retail counterparts. Another example is the "warehousing" function that is viewed by marketers as a "pre-retail" stage of distribution. This function, however, is classified by the census as partly retail and partly wholesale, and precludes any combined analysis by type of business. These kinds of concerns may not be viewed as traditional issues of quality, but they do affect the quality of census data in the marginal sense of their "fitness for use" (Juran (1980)).

6.5. Exogenous events

I will not dwell excessively on the problems resulting from the Internal Revenue Service's failure to provide adequate resources for the coding of non-employer establishments in the 1982 Economic Census. It suffices to say that these problems illustrate the special difficulties encountered in maintaining statistical quality standards while at the same time dependent on systems designed for non-statistical purposes and essentially outside the statistics producer's control.

7. A Proposal for an Additional Component to be Included in the Census Error Model

In most of the foreign countries where Nielsen works, we have had to conduct our own censuses of all types of retail establishments, since

[3] A Nielsen type mass merchandiser is defined as a retail establishment:
- Having 10 000 square feet or more selling area.
- Handling at least three major merchandise lines with no one line accounting for more than 80 % of the total selling space.
- Presenting a discount, high volume, fast turnover image through its advertising and promotion.

official census data are generally unavailable or are of uncertain quality. These are typically large-scale sample surveys rather than complete enumerations. Often they involve both a list frame for major establishments and an area sample frame for smaller establishments. For reasons that I shall explain later, these censuses are not taken at quinquennial intervals. They are spread over a cycle of several years so that annual estimates are produced throughout the cycle.

Similarly, in the United States our annual mass merchandiser "censuses" are not complete enumerations. Instead, we canvass, through actual field visits, all potential new establishments identified on a variety of lists. We also canvass establishments that are likely to have undergone changes in ownership or size, or are likely to be reclassified as in-scope or out-of-scope. All likely establishment "deaths" are verified by telephone call or field visit if necessary. All remaining establishments are visited on a sample basis, generally at a rate of about 1 in 6. We schedule around 1 800 field visits a year to maintain our universe file of around 7 000 in-scope mass merchandisers establishments.

Why do we conduct a partial census every year rather than a full census every five years? One reason is to schedule work more efficiently and to minimize the peaks and valleys in our survey workload. The more important reason, however, stems from the recognition of the way the data are to be used. That recognition leads to consideration of an important aspect of survey data quality – namely, how up-to-date the survey results are.

In considering the appropriateness of any census or survey error model, we need to take into account all components of error that affect the user of the data.

The Bureau of the Census makes use of a mean-square-error evaluation model that can be expressed (see Bailar and Kalleck (1980)) as

$$\mathrm{MSE}(x) =$$

$$\frac{\sigma_s^2}{kn} + \frac{\sigma_\mathrm{r}^2}{kn}[1 + (n+1\varrho] + \frac{2\,(n-1)}{kn}\sigma^{rs} + B^2\,,$$

in which the first term denotes the sampling variance, the second term the measurement or response variance and its correlated component, the third term the covariance between measurement and sampling deviations, and the fourth term, the square of the bias.

I suggest that we may need to extend the bias term beyond its usual definition to include an additional source of "error" that is critical to the census user but seldom considered. This component of "error" arises from the user's trying to infer the value of some variate X at time t_1 from its reported value at time t_0, the point at which the census was taken. With the exception of politicians and historians, few who make use of population census data are particularly interested in knowing what the population counts or characteristics were on April 1, 1980. On the contrary, users want to know what these counts or characteristics are at the time they are using the data.

We should recognize that the average "age" (defined as the time elapsed since the date at which the census was taken) of population census data over an entire decade of use is nearly seven years, ranging from a minimum of roughly nine months for state population counts to a maximum of over 12 years for sample characteristic data at, say, the postal code level just prior to the release of the following decennial censuses. Similarly, the average "age" of economic census data is nearly four years, ranging from a minimum of one year for state totals to a maximum of over seven years for some kind-of-business detail at the county level just prior to the release of the subsequent quinquennial business census.

If X is the "true" value of the variable of interest and $\hat{X}$ is the expectation of the published census value reported for that variate, the bias that affects the user at a point in time t can be thought of as

$$\widetilde{B} = X_t - \hat{X}_0 = (X_t - X_0) + (X_0 - \hat{X}_0),$$

where the second term is the usual bias term B.

Most of the literature on census quality ignores the first term in the above expression. Yet the difference $X_t - X_0$ may frequently exceed other sources of error, such as coverage error, content error, response error, and sampling error, which have been well documented in the literature.

It seems to us that this time-dependent "error" might, for varying lengths of time from the census reference date, be reasonably well estimated across sets of small geographic domains such as states, counties, places, tracts, enumeration districts, postal codes, or blocks. If this effort were successful, such estimates would help the user better understand the total "error" resulting from his use of obsolescent census data.

The following is an example of how such an evaluation might proceed.

Consider the task of estimating grocery store sales by state for 1982. In the absence of 1982 Economic Census data (or while awaiting publication of the 1982 Economic Census), one might look to the 1977 Economic Census results. Obviously inflation and population growth have moved sales forward, but let us assume that the user can make a good estimate of grocery store sales at the national level for 1982, using, for example, the Census Bureau's Monthly or Annual Retail Trade Survey. The question then is to what extent the 1977 state proportions of the U.S. total are still valid for 1982. The 1982 Census makes it possible to answer this question. That is, we can then examine:

$$\frac{X_j,\ 1982}{\sum_{j=1}^{51} X_j,\ 1982} - \frac{\hat{X}_j,\ 1982}{\sum_{j=1}^{51} \hat{X}_j,\ 1982},$$

for 51 states (counting the District of Columbia). We define

$$\frac{\hat{X}_j,\ 1982}{\Sigma\ \hat{X}_j,\ 1982} =. \frac{X_j,\ 1977}{\Sigma\ X_j,\ 1977},$$

so that the "error" is simply the difference in proportions between census years.

For this purpose, assume that we want a loss function defined in terms of average relative error, without regard to sign, at the state level, i.e.,

$$\frac{1}{51}\sum_{j=1}^{51}\left(\left[\frac{X_j,\ 1982}{\sum_{j=1}^{51} X_j,\ 1982} - \frac{X_j,\ 1977}{\sum_{j=1}^{51} X_j,\ 1977}\right]\Bigg|\frac{X_j,\ 1982}{\sum_{j=1}^{51} X_j,\ 1982}\right).$$

We may prefer to examine the root-mean-square of these relative errors.

The results of this exercise are shown in Table 2. They show that the 1977 state proportions of total U.S. grocery sales[4] differ from the 1982 proportions by an average of ± 7.7 % on a relative basis, and ± 9.7 % on a relative root-mean-square basis. The individual – state relative deviations are not quite symmetrical and the tails of the distribution are slightly thicker than a normal distribution's. They do not, however, differ from normality as much as one might expect. It is interesting to note, also, that the relative deviations are largely unrelated to the sales

[4] All grocery store figures used in this analysis are published by U.S Census of Business data for grocery stores without payroll, except for 1982 Current Retail Grocery Store Sales Estimates which are based on unpublished data provided to A.C. Nielsen Company. 1982 Economic Census results are preliminary.

Table 2. Accuracy of Alternative Estimators of Grocery Store Sales for 1982 in Percent (Grocery Stores Without Payroll)

	Percent of Total 1982 Census Grocery Sales	Average relative estimation error resulting from use of			
		1977 Census Proportions		1982 Current Survey	
State		Actual	Adjusted by population change	Theoretical*	Actual
California	11.40	− 2.7	+ 1.7	± 3.8	+ 2.6
Texas	7.85	−19.6	−11.8	± 5.8	− 3.3
New York	6.67	+10.6	+ 2.5	± 5.6	− 0.3
Florida	5.11	−12.5	− 3.3	± 6.7	− 6.0
Pennsylvania	4.72	+ 7.8	+ 1.4	± 5.1	− 5.0
Ohio	4.59	+ 9.6	+ 4.1	± 5.2	− 3.2
Illinois	4.29	+13.1	+ 7.4	± 6.3	+ 5.8
Michigan	3.46	+19.7	+12.9	± 6.7	+10.8
New Jersey	3.29	+ 7.7	+ 3.5	±10.3	+ 6.8
North Carolina	2.56	− 5.4	− 4.7	± 7.3	−13.5
Virginia	2.44	− 1.6	− 1.7	± 7.7	+24.3
Massachusetts	2.33	+11.7	+ 6.0	±10.5	− 5.1
Georgia	2.33	− 5.4	− 2.6	±10.8	−10.9
Indiana	2.19	+ 8.9	+ 4.6	± 9.4	+ 0.9
Louisiana	2.10	− 7.0	− 3.9	± 9.9	− 9.1
Washington	2.03	− 7.4	− 1.1	± 9.0	+ 4.2
Missouri	2.00	+ 6.8	+ 3.1	±10.0	+11.0
Tennessee	1.94	− 1.6	− 1.3	± 8.3	− 5.7
Maryland	1.87	+ 9.0	+ 5.5	± 8.3	− 8.6
Wisconsin	1.87	+ 5.3	+ 2.6	±13.9	−13.2
Oklahoma	1.61	−15.2	− 9.7	±11.6	−26.1
Minnesota	1.59	− 0.2	+ 2.0	±13.0	−11.2
Colorado	1.58	−16.1	− 9.9	±14.1	+11.9
Alabama	1.55	− 1.8	− 3.1	±13.2	− 0.7
Kentucky	1.50	+ 2.6	+ 0.3	±14.2	+ 3.0
Arizona	1.44	−12.4	+10.4	±10.8	− 2.4
Connecticut	1.41	+ 7.0	+ 3.0	±15.0	−13.7
South Carolina	1.33	− 2.8	− 0.4	±11.6	+ 2.2
Iowa	1.25	+ 3.2	− 2.6	±15.0	−14.0
Oregon	1.13	+ 4.6	+ 8.3	±10.2	+ 9.7
Kansas	1.01	+ 0.2	− 1.4	±20.6	−11.2
Mississippi	0.98	− 3.4	− 4.5	±17.4	+54.2
Arkansas	0.88	+ 3.3	+ 2.3	±22.7	+41.0
West Virginia	0.85	+ 3.8	+ 1.3	±14.0	− 8.0
Utah	0.64	−17.9	− 5.6	±21.6	−11.4
New Mexico	0.64	−13.0	− 8.3	±20.1	−23.0
Nebraska	0.60	+ 5.9	+ 2.2	±25.5	+30.4
Miami	0.53	+ 4.6	+ 2.2	±20.6	+15.5
New Hampshire	0.52	+ 1.5	+ 3.8	±23.2	+19.2
Nevada	0.51	−22.4	− 8.7	±19.0	−23.3
Hawaii	0.45	− 7.5	− 3.3	±20.6	−36.3
Idaho	0.44	− 2.2	+ 2.2	±21.6	+ 6.6
Montana	0.40	− 0.8	− 1.7	±18.4	−24.8
Rhode Island	0.35	+10.7	+ 4.0	±21.9	+53.3
Alaska	0.30	− 5.0	− 5.0	±20.1	−24.5
Wyoming	0.27	−21.3	− 9.3	±23.4	− 5.9
Delaware	0.27	+11.0	+ 6.5	±17.7	+ 6.4
Vermont	0.25	− 1.6	− 1.5	±24.2	+62.9
South Dakota	0.25	+ 4.1	− 1.4	±33.4	−29.4
North Dakota	0.23	+ 1.8	− 0.6	±38.7	+40.5
D. of Columbia	0.20	+12.1	− 8.3	±10.6	− 7.4
Average without regard to sign		7.7	4.1	14.6	15.5
Root-mean-square		9.7	5.2	18.3	21.1

* Equal to 80% of coefficients of variation

size of the state; that is, the "naive" assumption that nothing has changed since 1977 is no worse for small states than for large ones.

Two questions come immediately to mind:

i) Is this a unique set of results, or is it true of other time intervals as well?

ii) Are these "errors" a predictable function of time? What can be said about them for periods other than five years?

A straightforward calculation shows the following.

Average relative error (without regard to sign) in estimating individual state proportions of U.S. grocery sales from prior census years

Year being estimated	Year being used to make the estimate in percent		
	1967	1972	1977
1982	20.8	14.6	7.7
1977	14.3	7.8	
1972	8.2		

The errors are fairly consistent across equal-length time intervals, slightly less than linearly related to time (but not dramatically so), and almost normally distributed. It might also be possible to make a more rigorous determination of the size of the error as a function of time. We can compare the proportions, by state, reported by the Census Bureau's Annual Retail Trade Survey estimates each year with the prior year's economic census results. To do this we square the resulting differences, subtract the sampling variance component, and obtain the square of bias term as a residual. This will work, however, only if the time-dependent bias is large relative to the variance and if all other bias components are very small. Neither condition is likely to be fully met in this example. The exercise might be interesting nonetheless, especially for large states or perhaps for census regions or divisions.

Can the user improve on this 7.7 % average error by employing other sources of information? We have tested several possible estimators, and their performance is given in the table on the following page.

Note that, in Tables 2 and 3 (on pages 541 and 544), the individual state errors, expected and observed, are sample-based and vary inversely with the square root of their importance for sales. In contrast, the state estimators using prior economic census information, as well as those modified by the subsequent change in population weights, are largely unaffected by state sales importance. This suggests that it is better to use current census sample estimates for very large states (and census regions and divisions). Synthetic or census-based estimators are preferred for smaller states, at least for time intervals of five years. At other time intervals, different cut-off rules might apply. In general, the current retail trade estimates should be of relatively constant quality through time.

Note also that the combined synthetic estimator (#9), using equal (not optimal) weights, outperforms all the others. It can be further improved by an optimal choice of weights, and still further improved by use of a Stein-type estimator that also makes use of the Current Retail Trade Survey estimates. Such a procedure is used by Nielsen to produce individual county (as well as state) estimates of grocery store sales on an annual basis.

What about smaller geographic domains? Individual county results from the 1982 Economic Census are not yet fully released, so we had to go back to the 1977 Census results. Here we found the following (excluding 20 counties with zero or negligible sales).

How closely could one have estimated 1982 Grocery Store Sales Proportions by State, given knowledge of Total U.S. Grocery Store Sales for 1982?

	Estimation method	Average relative estimation error by state (percent)
1)	Use 1977 Census Grocery Sales Proportions	+ 7.7
2)	Continue 1972–77 Census Trend in Grocery Sales Proportions Forward to 1982	± 5.9
3)	Use 1982 Population Proportions	±10.5
4)	Use 1982 Personal Income Proportions	±14.4
5)	Use 1982 CBP Grocery Payroll Proportions	± 7.4
6)	Use 1977 Census Grocery Sales Proportions, Plus 1977–82 Change in Population Proportions	± 4.1
7)	Use 1977 Census Grocery Sales Proportions, Plus 1977–82 Change in Income Proportions	± 4.2
8)	Use 1977 Census Grocery Sales Proportions, Plus 1977–82 Change in CBP Grocery Payroll Proportions	± 4.7
9)	Use Linear Average of Methods 6, 7, and 8	± 2.8
10)	Use Census Current Retail Trade State Estimates of Grocery Sales for 1982:	
	– Actual Results	±15.5
	– Expected Results Due to Sampling Variability	±14.6

Average relative error in estimating 1977 Census Grocery Sales Proportions of Total U.S., by county, from 1972 Census Data (percent)

	Counties classified by 1977 Sales Importance (Terciles)		
All counties	Largest	Medium	Smallest
± 20.4	± 12.8	± 17.4	± 31.2

Analyses like the foregoing, applied to both demographic and economic census data at varying geographic levels, might help formulate an optimal or an improved allocation of resources between periodic censuses and interim current survey programs. These analyses might also help determine the level of geographic detail that should be included in a mid-decade census.

8. References

Bailar, B. and Kalleck, S. (1980): Evaluation of the 1977 Economic Censuses. Americal Statistical Association. Annual meeting, Houston.

Juran, J.M. (1980): Quality Planning and Analysis. McGraw Hill.

Morgenstern, O. (1963): On the Accuracy of Economic Observations. Princeton University Press, Princeton NJ.

Nielsen Metered Television Markets (1985): New York, Los Angeles, Chicago, San Francisco, Philadelphia, Detroit, Boston, Washington, D.C., Dallas.

U.S. Bureau of the Census (1978): An Overview of Population and Housing Census Evaluation Programs Conducted at the U.S. Bureau of the Census. (Preliminary draft for meeting of the Census Advisory Committee of the A.S.A.), p. 39 and p. 55.

U.S. Bureau of the Census (1984): Content Evaluation of the 1977 Economic Censuses. SRD Research Report number, Census/SRD/RR–84/29, p. 23.

TABLE 3

PLOT OF STANDARDIZED REGRESSION RESIDUALS WHERE:

Y = UNSIGNED PERCENTAGE DIFFERENCE BETWEEN 1982 CENSUS MONTHLY RETAIL TRADE ESTIMATE OF GROCERY STORE SALES (BY STATE) AND ACTUAL 1982 CENSUS GROCERY STORE SALES (RE82CSA), AND

X = INVERSE SQUARE ROOT OF STATE GROCERY STORE SALES (PRED)

(STATES SEQUENCED IN DESCENDING ORDER BY SALES SIZE)

```
       -3.0          0.0           3.0
SEQNUM  O:.............:.............:O     RE82CSA       *PRED      *RESID
     1  .             *.              .      2.6000      4.4326     -1.8326
     2  .             *.              .      3.3000      5.3417     -2.0417
     3  .            * .              .      0.3000      5.7949     -5.4949
     4  .              *              .      6.0000      6.6206     -0.6206
     5  .             *.              .      5.0000      6.8887     -1.8887
     6  .            * .              .      3.2000      6.9856     -3.7856
     7  .             *.              .      5.8000      7.2257     -1.4257
     8  .              .*             .     10.8000      8.0459      2.7541
     9  .             *.              .      6.8000      8.2511     -1.4511
    10  .              . *            .     13.5000      9.3538      4.1462
    11  .              .     *        .     24.3000      9.5811     14.7189
    12  .            * .              .      5.1000      9.8047     -4.7047
    13  .              *              .     10.9000      9.8047      1.0953
    14  .          *   .              .      0.9000     10.1132     -9.2132
    15  .              *              .      9.1000     10.3276     -1.2276
    16  .           *  .              .      4.2000     10.5042     -6.3042
    17  .              *              .     11.0000     10.5827      0.4175
    18  .            * .              .      5.7000     10.7451     -5.0451
    19  .             *.              .      8.6000     10.9443     -2.3443
    20  .              .*             .     13.2000     10.9443      2.2557
    21  .              .     *        .     26.1000     11.7950     14.3050
    22  .              *              .     11.2000     11.8689     -0.6689
    23  .              *              .     11.9000     11.9064     -0.0064
    24  .         *    .              .      0.7000     12.0211    -11.3211
    25  .          *   .              .      3.0000     12.2198     -9.2198
    26  .          *   .              .      2.4000     12.4718    -10.0718
    27  .              *              .     13.7000     12.6038      1.0962
    28  .          *   .              .      2.2000     12.9773    -10.7773
    29  .              *              .     14.0000     13.3861      0.6139
    30  .            * .              .      9.7000     14.0790     -4.3790
    31  .             *.              .     11.2000     14.8919     -3.6919
    32  .              .              *     54.2000     15.1181     39.0819
    33  .              .        *     .     41.0000     15.9540     25.0460
    34  .           *  .              .      8.0000     16.2331     -8.2331
    35  .           *  .              .     11.4000     18.7077     -7.3077
    36  .              . *            .     23.0000     18.7077      4.2923
    37  .              .   *          .     30.4000     19.3212     11.0788
    38  .            * .              .     15.5000     20.5576     -5.0576
    39  .             *.              .     19.2000     20.7543     -1.5543
    40  .              .*             .     23.3000     20.9568      2.3432
    41  .              .     *        .     36.3000     22.3102     13.9898
    42  .        *     .              .      6.6000     22.5623    -15.9623
    43  .              *              .     24.8000     23.6636      1.1364
    44  .              .         *    .     53.3000     25.2974     28.0026
    45  .             *.              .     24.5000     27.3243     -2.8243
    46  .      *       .              .      5.9000     28.8024    -22.9024
    47  .      *       .              .      6.4000     28.8024    -22.4024
    48  .              .           *  .     62.9000     29.9323     32.9677
    49  .              *              .     29.4000     29.9323     -0.5323
    50  .              .   *          .     40.5000     31.2066      9.2934
    51  .     *        .              .      7.4000     33.4653    -26.0653
SEQNUM  O:.............:.............:O     RE82CSA       *PRED      *RESID
```

Received May 1986
Revised December 1986

Journal of Official Statistics
Vol. 2, No. 4, 1986, pp. 545–553
© Statistics Sweden

How Increased Automation Will Improve the 1990 Census of Population and Housing of the United States

Peter Bounpane[1]

Abstract: The U.S. Bureau of the Census will increase significantly the automation of operations for the 1990 Census of Population and Housing, thus eliminating or reducing many of the labor-intensive clerical operations of past censuses and contributing to the speedier release of data products. An automated address control file will permit the computer to monitor the enumeration status of an address. The automated address file will also make it possible to begin electronic data processing concurrently with data collection, and, thus, 5–7 months earlier than for the 1980 Census. An automated geographic support system will assure consistency between various census geographic products, and computer-generated maps will be possible. Other areas where automation will be introduced or increased are questionnaire editing, coding of written entries on questionnaires, and reporting of progress and cost by field offices.

Key words: 1990 U.S. Census of Population and Housing; increased automation; automated address control file; automated geographic support system; earlier processing.

1. Introduction

The U.S. Census Bureau began planning the 1990 Census of Population and Housing – the Bicentennial Census of the United States – several years ago. Even though April 1, 1990, is still 3 years away, an early start was necessary because of the complexity of the issues and the time needed to implement decisions. The broad range of issues addressed in census planning are described in Bounpane (1985). Our goals for 1990 are to publish more timely data products and to make the whole census process more cost-effective while at the same time maintaining a high level of accuracy. In other words, we are attempting to make the census process more productive. We hope to achieve greater productivity by automating outmoded clerical operations and by entirely rethinking the data collection and data processing stages of the census.

Over the last century, the census has played an important role in the history of automated data processing in the United States. By 1890, the U.S. census had become an encyclopedic enumeration of the American people. The 1890 Census marked a great increase over previous censuses both in the number of inquiries and the volume of data tabulated and published. Census officials, realizing that something had to be done to speed up the processing and tabulation for the 1890 Census, gave a young engineer named Herman

[1] Assistant Director, Bureau of the Census, Washington, D.C. 20233, U.S.A.

Hollerith the assignment of constructing a quicker tabulating device. The electromechanical tabulating machine Hollerith developed for the 1890 Census – which read punched cards by electrical pulses – revolutionized both census-taking and statistical tabulating. Hollerith's machine was soon used worldwide in business and census applications (Austrian (1982)).

Hollerith's invention allowed greater volumes of data to be processed, more sophisticated cross-classifications, and all in a shorter time and at less cost. His punch-card system was modified and improved by the census machine shop for each successive census over the next 60 years.

Eventually, computers replaced the tabulating machine for processing data, and the census was again at the forefront of the technological revolution. UNIVAC-1, the first major computer system for civilian use, was installed at the Census Bureau in 1951 and was used to process part of the 1950 Census. Though large, cumbersome, and slow by today's standards, UNIVAC-1 was a major advance from the Hollerith tabulating system. Computers were used to process all of the 1960 Census, and, of course, the 1970 and 1980 Censuses.

Another new device accompanied the 1960 Census: FOSDIC, a replacement for keying, was introduced for entering data into the computer. FOSDIC is an acronym for Film Optical Sensing Device for Input to Computer. Questionnaires were microfilmed by special page-turning cameras, and FOSDIC read the data from microfilm into the computer. This advance, which was developed to meet the specialized needs of the decennial census, eliminated the need for key-punchers, saved time, and improved quality. FOSDIC has been used in the last three censuses.

The point of this brief history is that the decennial census, because of its massive workload and unique character, has called forth new technology, new tabulating, computing, and automated equipment to speed up the processing of census data.

As we examined our experience from the 1980 Census, we found that while the census was generally a success there was need for improvement. We determined that much of the improvement in timeliness, accuracy, and cost-efficiency could come from taking a fresh look at automation and increasing automation in the census.

While we do not yet know whether a specific automation decision will save money, we believe that our decisions will lead to a more efficient and accurate census. We will invest in automation that could reduce costs or that is necessary for maintaining or improving the quality of the census. Automating census operations will allow us to replace labor intensive and error-prone clerical operations with automated techniques that are quicker, more accurate, and easier to control.

While the automation advances we plan for the 1990 Census will not involve the development of new technologies, they will be based on innovative applications and refinements of existing technologies. The Census Bureau has embarked on a vigorous program to examine automation alternatives in test censuses before making choices for the 1990 Census. Since we are contemplating significant changes in automation for 1990, I will first describe how the 1980 Census was taken so the departures will be more easily understood.

2. 1980 Census

The 1980 Census was taken using the mail-out/mail-back procedure in areas of the country that contained 95 percent of the population. We purchased address lists for some of these areas and listed addresses ourselves elsewhere. In all cases, the address lists were then checked and updated by the U.S. Postal Service and our own field personnel. The USPS delivered questionnaires to each housing unit a few days before census day and householders were asked to fill them out and mail them back to a temporary census district office on April 1st. The aim of this approach was to complete as much of the census as possible by the less costly mail method and then to do the costly and time-consuming follow-up of those housing units that did not return a questionnaire. We had received questionnaires for about 83 percent of the households within 2 weeks of our initial mailing. A large work force (270 000 at peak) visited nonresponding housing units and vacant units. In sparsely populated areas where mail-census procedures were not suitable, census enumerators went door-to-door to take the census (Cho and Hearn (1984, pp. 241–263)).

We set up 409 temporary district offices to carry out data collection. Most of the operations were done manually. For each office, a large number of clerks were hired to make changes (additions, deletions, corrections) to the address lists, check in mail-returned questionnaires and edit the questionnaires for completeness and consistency, assign housing units for follow-up, monitor the enumeration of the nonresponding units, and tally preliminary counts. Many of these operations can be considered "processing," but processing did not begin in earnest until the collection offices completed their work, closed, and shipped their questionnaires to one of three processing centers. The offices generally closed 5–7 months after census day.

At the processing centers, the questionnaires were microfilmed and the data read into the computer by FOSDIC. Though processing center operations were largely automated, written entries for many questionnaire items (e.g., ancestry and occupation) were manually given numeric codes prior to computer processing.

This system worked very well considering the amount of manual work involved and the sharp division between data collection and data processing. First, the Census Bureau met the deadline dictated by law for the release of apportionment and redistricting counts. Apportionment is the process whereby a state is awarded a share of the 435 seats in the House of Representatives based on its population; redistricting refers to the process of redrawing the boundaries of legislative districts within states based on the principle of "one person/one vote." Second, many of the small-area data were issued earlier than for the previous census. For example, the 1980 Census data for 2.5 million blocks used in redistricting were produced in less than 12 months. For the previous 1970 Census, similar data for 1.7 million blocks took 18 months to produce. Third, many more data, especially for race and Spanish-origin groups, were published. Still, we did not release some of the data products, particularly those based on the sample questions, as quickly as planned. (This delay was due in part to budget problems that forced us to cut staff and temporarily suspend sample coding operations.)

For the 1990 Census, we want again to meet our deadlines and we want to release other data products more quickly than before, as well as keep costs reasonable and make the counts as accurate as possible.

3. Automation Plans for 1990

We have identified a number of areas that are candidates for automation, and have already begun to test some of them.

3.1. *Geography*

Geographic materials are essential to a successful census for two reasons: First, having correct and legible maps helps our enumerators find every housing unit so that we have a complete count; and second, having correct boundaries and geographic information helps us assign each housing unit and the people who live there to the appropriate land area. One of our problems in the 1980 Census was that our geographic materials, including the maps, were produced in separate operations involving a great deal of clerical work. This process was slow and error-prone, leading to delays in production and inconsistencies in some of the products.

For 1990 we are automating our geographic support system, which we are calling TIGER (Topologically Integrated Geographic Encoding and Referencing system). TIGER will integrate all the geographic information that was produced in separate operations in 1980. This will allow us to produce the geographic products and services for 1990 from one consistent data base, and will help us avoid some of the 1980 Census delays and inaccuracies. Having computer generated maps that match the geographic areas in our tabulations will be an improvement over the clerical operations of the 1980 and earlier censuses. For a full discussion of the automated geographic support system, see Marx (1986).

3.2. *Address control file*

Another improvement planned for the 1990 Census is the development of an automated address control file. Since we will again use the mail-out/mail-back methodology, an accurate and up-to-date address control file is essential. In 1980, although the initial control list of addresses was computerized, changes in the address file during the census were made manually. For 1990, we will have continuous access to the automated address control file so that we can keep the list current.

With an automated address file, it will be much easier to determine whether or not we included a specific address in the file. It also will be possible to update the file where we missed an address in earlier operations. We can imprint bar codes (like those on supermarket items) on the questionnaires and use electronic equipment to read the information in the bar codes. Thus, we can use the computer for checking in and keeping track of census questionnaires instead of doing check-in manually as we did in 1980. As a result, it will be easier for our enumeration staff to identify nonrespondents' addresses.

Finally, with an automated address list, we can update the list and use it in future Census Bureau operations. In our 1985 and 1986 test censuses, we successfully implemented an automated address control file and automated check-in.

3.3. *Earlier data conversion*

One of the most promising ways to take advantage of automation in the census, and our biggest challenge, is to convert the data on the questionnaires into a computer-readable format earlier in the census process than in past censuses. This approach is essential if we are going to take full advantage of automation and release data products quicker.

In 1980, the conversion of data to machine-readable form did not begin until after the district offices completed all enumeration, edits, and follow-ups and shipped all questionnaires to one of the three automated processing centers. This was a sequential process. This meant that many completed questionnaires that could have been automatically processed early in the census, lay around for several months until the district offices closed. Also, because we did not have an automated address control file, we had to process all the questionnaires for an enumeration district in one batch.

The automated address control file for the 1990 Census will allow us to conduct flow processing, and to do it concurrently with data collection. An earlier start in 1990 (5–7 months ahead of the 1980 schedule) will allow more time for review and correction and will enable the computer to assist in certain census operations. It will contribute to the early identification of enumeration problems. Also, by converting questionnaire data to machine-readable form sooner, we can minimize the loss of data when original questionnaires are accidently damaged or destroyed. Finally, and perhaps most importantly, it will help us meet our goal of disseminating data products more quickly.

Planning for concurrent processing in the 1990 Census has centered on two major questions: Where and how would it be done? The "where" issue involves the number of processing offices and the degree of centralization or decentralization. In 1980 we processed the census questionnaires sequentially and had three processing centers. With concurrent processing, having so few centers probably would not be feasible because of the need to move materials quickly between processing and collection offices. Greater centralization of processing activities also places greater staffing burdens on the center, i.e., the need to hire more employees in one area.

We weighed these concerns against problems related to decentralization – the need for more hardware and the difficulties of controlling and supporting many processing offices.

The "how" issue involves the technology we will use to convert questionnaire data into a computer-readable format. In the 1980 Census, we employed the FACT-80 system (with FOSDIC technology as the base) to convert microfilm directly to computer tape. FACT is an acronym for FOSDIC and Automated Camera Technology. The complete data-conversion system consists of high-speed cameras that film the questionnaires, film developers to process the rolls of microfilm, and the FOSDIC machines that read the data from microfilm to computer tape.

We also looked at key-entry as a primary data conversion methodology. Both FOSDIC and keying are tested methodologies that have proved workable over the years. Because there are technical limitations to how many FOSDIC systems we can build and maintain for 1990, we had considered data keying to give us maximum flexibility in decentralization. Keying was not considered as a viable option as the sole data conversion technology for the entire census because of the large numbers of keyers and key stations that would be required.

Earlier in our planning we had also considered a third technology – optical mark recognition (OMR). OMR provides direct input of data into the computer, whereas with FOSDIC the questionnaires must be filmed first. As with keying, we considered OMR to allow us more flexibility in decentralizing our processing. We tested OMR in our 1985 Census in Tampa, Florida. Based on some of the problems experienced with OMR in this test, and on other concerns about cost, timing, environmental controls, and so on, we decided not to pursue further testing of OMR technology for use in 1990. We will, however, consider testing OMR and other technologies in 1990 for possible use in the 2000 Census.

In April 1986, after reviewing these two main issues at planning conferences and in internal working groups, we were able to reach some decisions. We have decided to set up eleven processing centers for the 1990 Census where we will use FACT 90 (an update of the 1980 system, still with FOSDIC as the base) to convert the data to machine-readable format.

We determined that having two primary data conversion technologies (FOSDIC and keying) would have excessively complicated our processing system for 1990. We will use keying only as a supplement to FOSDIC for entering some of the handwritten data on the questionnaires into computer-readable form.

We will have two types of district offices for which the questionnaire flows will be different. For district offices in certain high population density areas the processing centers will receive the questionnaires, perform automated check-in using laser sorters, immediately convert the questionnaires to computer-readable form, and thereby perform an automated review (edit) of the questionnaires. The district offices covered by these processing offices will likely correspond to some of our "centralized" offices in 1980 – the more hard-to-enumerate urban cores where recruiting enough temporary census workers can be difficult. These district offices will not need to hire many office clerical workers and can concentrate on field follow-up activities for households that did not mail back their questionnaires or that mailed back incomplete questionnaires.

District offices in the rest of the country will receive the returned questionnaires; use pencil-shaped, electronic "wands" attached to micro computers to read the bar codes on the questionnaries and, thus, perform automated check-in; and conduct clerical edits for completeness. Once questionnaires pass the edit, they will be sent on a flow basis to a processing center for data conversion (using FACT-90).

This decision represents a careful balance of staffing, equipment, and workload considerations as they relate to the processing and collection offices. We will have an automated address control file and automated check-in for the entire area covered by the mail-out/mail-back census, and we will achieve our goal of concurrent processing by converting questionnaire data to computer-readable format on a flow basis, several months earlier than for the 1980 Census.

So far I have discussed our plans with regard to automating geographic materials and the address control file and beginning data conversion earlier. We will increase or improve automation in other areas to help speed up the census and make it more accurate, and I will discuss briefly a few of these areas.

3.4. *Computer edits*

One area is questionnaire edit. Edit is a repetitive and monotonous job better suited for computers than people. Entering data from the questionnaires to the computer earlier in the census process will allow computer editing of the questionnaire data earlier than ever before. These edits will check the completeness and consistency of the data. In 1980, the questionnaires were manually edited in the district offices, basically to check that they had been answered completely; then, once the questionnaires went through the FOSDIC machines, the computer edited them for completeness and consistency. For 1990, manual editing would be eliminated in some district offices and replaced by computer edits.

3.5. *Automated coding*

Another promising automation technique relates to the coding of handwritten entries on the questionnaire. In 1980, manually coding the handwritten entries on questionnaires involved a large, time-consuming, and costly clerical operation. For 1990, we might be able to key handwritten responses into the computer and develop software that would assign the appropriate computer-readable codes. We cannot eliminate all clerical involvement in coding, because some handwritten responses will be incomplete or uncodable and will have to be handled by our referral units. We will, however, be able to significantly reduce the amount of manual work and, thus, save time and improve the quality of the data. Instead of a clerk having to look up the occupation "statistician" in a reference manual, find the numerical code, and fill the appropriate coding box on the questionnaire, the clerk can type in the word "statistician" and the computer will automatically assign a code and enter that onto a computer record. Thus, the time-consuming looking-up and circle-filling are eliminated. At this time, we do not know precisely the extent that the computer will be able to assign codes without clerical intervention.

3.6. Management and administration

We will also use automation to help us plan and monitor the census. The Census Bureau is developing an elaborate automated management information data base to see that we meet important dates in making decisions for the 1990 Census. The management information system was used to help us keep track of operations for our 1985 and 1986 test censuses. In addition to serving as an aid in planning the 1990 Census, the management information system will give us up-to-the-minute cost and progress data so that we can monitor actual 1990 Census operations. In 1980, cost and progress reports were not integrated with other management reports, and some of the cost and progress information was several days old by the time managers received it.

Automation will help us control and monitor many other administrative functions. We will have an automated payroll system, as in 1980. And for 1990, we will also have, on a microcomputer, a new automated employee file that will help us organize needed information about our large temporary work force. (We did this in our 1985 test census.) For instance, we will know whether we are meeting our hiring goals in each enumeration area and we can use the file to help us make enumerator assignments. We will also have a new automated inventory control system to manage the procurement and distribution of the large volume of specialized supplies needed to take the census.

3.7. Data products

Finally, we are looking at further automation of our tabulation and publication operations for the 1990 Census. The actual tabulation of data was fully computerized for the 1980 Census, but the design and review of specifications and the review of test data was largely manual. We want to use the computer in our development of specifications and the analytical review of the tabulated data for 1990. This review, which looks for errors and anomalies in the data, is essential to maintaining the quality of our data products. Using the computer will improve this analysis.

New automation techniques will also play a part in the dissemination of our data products for the 1990 Census. While the Census Bureau will continue to produce paper reports and large summary computer tape files, we must also address the needs of small computer users who will want products on floppy disks. Another new development we will consider for 1990 will be an online data base in which users can access summary data from their office computers using the telephone. The Census Bureau has already implemented such a system, called CENDATA, on a limited basis. There may be other developments in the next few years – such as improvements in laser disks – that we will be able to take advantage of for the 1990 Census. Fortunately, our final decisions on tabulations and data products can be made later in the decade, so we can take advantage of new technologies.

4. Closing

There is a sense of excitement at the Census Bureau about these automation possibilities, but some words of caution should be added. The systems developed must be simple, because they will be operated by a temporary work force with minimal training. The systems must be fully tested, proven reliable, and essentially “fail safe” to avoid crippling breakdowns. The equipment must be reasonably priced and should either continue to have value to the Census Bureau or be marketable to someone else upon completion of the census.

Most of all, as we look to increasing automation in the census, we must take care to ensure that the confidentiality of the data we collect is maintained both in fact and in appearance. Only by maintaining the confidentiality of the census process can we ensure a high level of public trust and cooperation. The Census Bureau is proud of its record of protecting confidentiality and is constantly looking for ways to maintain and improve that protection.

The Census Bureau does not release data about individuals to anyone, including other Federal government agencies. But the sometimes menacing implications of technology require that we increase our efforts to convince individuals that they cannot be harmed by answering the census and that the information they provide is strictly confidential by law.

Automation is one of the key areas we are examining as we plan the 1990 U.S. Census of Population and Housing. There are many other issues, of course, that go into making a successful census: the basic procedures we will use to collect the data, the content of the questionnaires, hiring good temporary staff, and promotion of the census, including contacts and consultation with various groups and individuals interested in the census. However, automating many of the census tasks performed clerically in 1980 and previous censuses can help us to take the census more quickly, allowing us to meet our legal mandates for releasing apportionment and redistricting counts and to release other data products quickly. Automation could also help us introduce cost-efficiencies into many areas, improve accuracy, and also allow for better control of the census process.

Traditionally, U.S. census data collection and much of the census data processing (e.g., questionnaire check-in against the address control list, edit of questionnaires for completeness, and coding of handwritten responses) have been paper- and people-intensive tasks. The use of automated equipment can help to deal with the mountains of paper and the thousands of clerical tasks in a much more efficient and controlled way. Hiring, training, and finding space for all the people who have been needed to perform the numerous operations in past censuses have required a lot of time and money. While the 1990 Census will also require a large number of temporary workers, we are looking at ways to cut down on the number of labor-intensive activities and to use automated systems to control the census process.

We have been working on our automation plans for some time now. We tested some new approaches in our test censuses in 1985 in Tampa, Florida, and in Jersey City, New Jersey, and conducted further tests of automation this year in part of Los Angeles County, California, and in several counties in east central Mississippi. These tests are very important as laboratories where we can try out optional approaches. There will be further testing in 1987 and a dress rehearsal in 1988.

While there are many decisions yet to be made and problems to be worked out, we have progressed far enough in our automation planning to say this: there will be significantly more automation in the 1990 Census than in any previous census. We will make innovative use of automation techniques to perform data-entry earlier than ever before. We will have an automated geographic support system. We will edit questionnaires by computer. And we have already implemented an automated address control file, automated questionnaire check-in, and an automated management information system in our test censuses, and plan to have these features in 1990. Thus, we are optimistic that we are on the verge of important advances in applying automation to census-taking. That is fitting since 1990 will mark the 200th anniversary of the first U.S. census in 1790.

5. References

Austrian, G.D. (1982): Herman Hollerith: Forgotten Giant of Information Processing. Columbia University Press, New York.

Bounpane, P.A. (1985): Plans and Issues Facing the 1990 Decennial Census. Government Information Quarterly, Vol. 2, No. 4, pp. 369–387.

Cho, L.J. and Hearn, R.L., (eds.) (1984): Censuses of Asia and the Pacific: 1980 Round, pp. 241–263. East-West Population Institute, Honolulu.

Marx, R.W. (1986): The TIGER System: Automating the Geographic Structure of the United States Census. Government Publications Review, Vol. 13, pp. 181–201.

Received July 1986
Revised December 1986