Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 95–112

On Estimating Census Undercount in Small Areas

Cary T. Isaki, Linda K. Schultz, Gregg J. Diffendal, and Elizabeth T. Huang[1]

Abstract: Net undercount rates in the U.S. decennial census have been steadily declining over the last several censuses. Differential undercounts among race groups and geographic areas, however, appear to persist. In the following, we examine and compare several methodologies for providing small area estimates of census coverage by constructing artificial populations. Measures of performance are also introduced to assess the various small area estimates. Synthetic estimation in combination with regression modelling provide the best results over the methods considered. Sampling error effects are also simulated. The results form the basis for determining coverage evaluation survey small area estimates of the 1990 decennial census.

Key words: Census; undercount; adjustment; small area estimation; synthetic estimation; regression; artificial population; simulation.

1. Introduction

1.1. Background

While the net undercount of total U.S. population at the national level has been declining over the last four censuses (from 4.4% in 1950 to 1% in 1980), the 1980 census, according to the U.S. Census Bureau's post enumeration program (PEP), experienced differential net undercount for states that ranged from a 2% overcount to a 6% undercount and the net undercount by race at the national level was 0.1% overcount for White and other persons and 4% undercount for Black and Hispanic persons. The differential undercounts have been the basis for court cases requesting an adjustment of the 1980 census counts.

In the following we present and compare the performance of several methods of adjusting census population counts for small areas. We begin with a brief description of the sources of undercount measurement likely to be available for use. A discussion of assessment of adjustment methodologies and the particular approach used in our research is then described. Succeeding sections describe the methods of adjustment and adjustment results that have been reported previously, introduce recent adjustment results, and summarize our findings.

Subsequent to the submission of our paper it was announced (at an administrative level

[1] U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A. The views and conclusions contained herein are those of the authors and are not necessarily that of the Bureau of the Census.
Acknowledgements: We are grateful to three referees and the editor for their helpful comments that improved our presentation.

above that of the Bureau of the Census) that the 1990 decennial census would not be adjusted but that resources will be concentrated on producing the most accurate enumeration possible. Some of the reasons given were (1) adjustment would be controversial; techniques are available but there are questions about the validity of their results; there is no unique system generally accepted by statisticians, (2) adjustment would raise suspicions in the public mind about the reliability and integrity of the census; reduce their willingness to respond and, (3) adjusting the count may create the appearance of changing the numbers to achieve a desired political outcome.

The bureau will, however, conduct a post enumeration survey (PES) and produce demographic analysis (DA) estimates of total population. Hence small area estimates of total population under the PES program will be provided. Such estimates will not be timely for purposes of census adjustment, however, but will provide a basis for assessing the quality of the 1990 census counts at the small area level.

Despite the decision to not adjust the 1990 census, we believe that interest in the presentation of our results concerning possible methods of census adjustment remains (legislation has been proposed to adjust the 1990 census for undercount). First, if census adjustment is decided upon for a future census, a single method of adjustment of the census at the small area level will be needed. Second, the best of the competing adjustment methods can be fairly chosen when all are evaluated under similar conditions and against a common standard. Third, the level of magnitude of the errors in adjustment results can be compared and an assessment of their acceptability can be made by users of the data. Finally, small area estimates of undercount are required for the evaluation program and the reader could interpret the following results for that purpose.

1.2. Overview of undercount measurement

Adjustment of the census for undercount at the small area level assumes that estimates of the undercount for larger areas or categories are available. We term such estimates direct estimates. For example, if accurate estimates at the state level exist, then subunits of states, such as counties, are the small area of interest. The manner of estimation of county total population is then the small area adjustment problem to be addressed.

A description of the 1980 census post enumeration program, termed the 1980 PEP, can be found in Cowan and Bettin (1982). Three excellent sources on 1980 census adjustment issues, methods and critique are Citro and Cohen (1985), Ericksen and Kadane (1985), and Freedman and Navidi (1986), the latter two with discussion. It is assumed that the reader is familiar with the content of the latter two papers. A post enumeration survey (PES) is planned for evaluating the 1990 census and together with the demographic analysis (DA) estimates will constitute the evaluation program. The 1990 PES will again be based on a matching procedure but is expected to be an improvement over the 1980 procedure because (1) the sample design will be specified to support postenumeration purposes; the sampling unit will be the block and adjustment factors suitable for small area estimation will be produced, (2) computer matching algorithms (Jaro (1985) will enable faster matching results and quicker resolution of unmatched cases, and (3) a larger sample size of housing units is anticipated. The DA estimates will also be less reliant on the use of models.

For 1990, the Bureau of the Census has determined that demographic analysis and the PES will be used to estimate the net undercount. The results from the two sources are to be combined in a fashion that is yet to be determined. For present purposes, we essentially assume that the PES will be the main source of undercount estimates.

1.3. Recent research

Under a PES, dual system estimation (DSE) is the method used to directly estimate the population. DSE has direct parallels with capture-recapture estimation. (See Seber (1982) for a detailed exposition on capture-recapture methods.) However, DSE also has unique problems as well. For example, as Cowan and Malec (1986) point out, responses to the census and separately, to the PES, may be dependent due to household clustering. Or, from Jabine and Bershad (1970), responses to the census and the PES may be correlated; they could also be misclassified, matched or not matched in error, subject to imputation, etc. Ericksen and Kadane (1985) and Isaki (1986) have proposed dual system estimation type procedures that considerably reduce the bias over that of the DSE when knowledge of a lower bound on the response correlation parameters is available. Isaki and Schultz (1987a) have also shown that in the presence of matching error, the DSE performs better (less bias) than other dual system estimation procedures when response correlation is positive (response to the census is likely to result in a response to the PES as well). Cowan and Fay (1984) address the issues of imputation in the DSE as applied in the 1980 PEP. Wolter (1986) provides an excellent summary of the variety of type of coverage models underlying DSE.

2. Aims of Small Area Estimation for Adjustment

The ultimate goal of small area adjustment for census undercount is to produce an adjusted census figure for each and every figure produced in previous censuses. Only adjusted figures should be reported; hence adjustment is not visible to the user. In addition to adjusting census population counts down to the lowest publication level, namely blocks, the goal also implies adjustment of various other person type characteristics such as income, labor force status, etc., and household characteristics such as rent paid, number of rooms, etc. Our present goals are more modest. We are concerned with population count adjustment down to the block level. We are investigating the effects of several undercount adjustment methodologies to produce adjusted population counts at various geographic levels.

Our reasons for emphasizing population count adjustment are twofold. First, a complete adjustment of population counts and housing unit characteristics, in our opinion, requires joint estimation of undercount of both types of characteristics. At a minimum, it requires estimation of persons missed in counted households and estimation of persons missed in missed households. At present, even this type of estimation is not available in any proposed PES. Second, population count adjustment is possible in the short run and useful for many applications of census data. Consequently, emphasis is being placed on developing and studying the results of small area adjustment as they pertain to population counts. In a 1986 test of adjustment related operations, undercounted (and overcounted) persons at the census block level were placed in a special publication category (group quarters).

Much work needs to be done to achieve our modest goals of population count adjustment. Adjustment methods need careful study and evaluation. Part of that work is the focus of this paper.

3. Assessment

3.1. Standards

On the one hand we have the (traditional) results of the census. On the other hand we have the results of one or more adjustment method applications. The question of when adjustment is deemed superior to the census requires consideration. Several measures of

performance have been proposed to assess the results of adjustment and the census. Each measure of performance requires that the actual (true) population of each small area (a standard) is known. Given the limitations of both the 1980 demographic analysis and 1980 PEP estimates of undercount, an alternative standard was necessary for assessment. The standard we selected was the artificial population which we used to simulate the undercount.

The artificial populations (we used several in our work) were constructed using census substitutions as a proxy for persons missed in the census. The variable, census substitutions, is a count of the number of persons imputed into housing units. Persons were imputed into housing units when no form was completed but people may have lived in the housing unit, when we knew only the number of people living in the unit, for machine failure (census forms destroyed) or when field counts for an area (enumeration district (ED), averaging 800 persons) were larger than the processed counts. In our opinion census substitutions reflected an indication of difficulty of enumerations as well as a clustering of error (machine failure in processing a batch of documents from the same area) and hence a reasonable proxy for undercount. Furthermore, preliminary analysis using 1980 PEP state data indicated that the census substitution rate was the most important explanatory variable of several types of nonmatch rates in the PEP. The census substitution rate was also one of two explanatory variables selected in a regression of district office (DO) 1980 PEP nonmatch rates. (A district office is a census administrative area covering approximately one half million persons.) The nonmatch rate in the PEP is by definition the ratio of estimated total number of persons in the PEP not matched in the census to the PEP estimated total number of persons. The nonmatch rate estimates the miss rate of the census (under ideal conditions). Finally, a census population count file at the enumeration district level, the lowest level of geography available to us at the time, contained about twenty variables (including census substitutions). When we imposed the requirement of age-race-sex categories (we anticipated use of an adjustment scheme requiring such information) we were left with a handful of variables. Fortunately, census substitutions were available by age-race-sex.

We constructed three artificial populations labelled AP1, AP2 and AP3 to be used as standards. The three artificial populations constructed by five age categories (0–14, 15–29, 30–44, 45–64, 65 and over) by sex by three race categories (Black, non-Black Hispanic and Rest) at the ED level are exhibited in the figure below.

Population	Enumerated part	Undercount
AP1	Census – substitution	Substitution
AP2	Census	$F_{DA1} \times$ substitution
AP3	Census	$F_{DA2} \times$ substitution

AP1 treats the 1980 census count as the standard and the census count minus substitution as the resulting enumeration. Because the resulting enumeration undercount of AP1 (for example, at the national level by race) was low for Blacks when compared to 1980 demographic analysis (estimates including an assumed 3.5 million illegal aliens) we constructed AP2 and AP3. DA does not provide for a non-Black Hispanic estimate and hence the undercount parts of AP2 and AP3 for the non-Black race groups differ in the way we

partition the DA estimate. We focus on the Black group first. Let N_{DA} denote the DA estimate (by age-sex) of total Black population. Then F_{DA} = (N_{DA} – enumerated part)/ substitution is computed for each age-sex category. The same factor is used in constructing AP2 and AP3 for the Black categories. For Blacks under AP2 and AP3 then, N_{DA} represents the standard at the ten (age-sex) U.S. total population levels.

For non-Black Hispanics under AP3, the Black F_{DA}'s were used. The resulting AP3 non-Black Hispanic totals were subtracted from the DA estimated non-Black totals to give the resulting N_{DA} for the Rest category. The F_{DA}'s for the Rest are defined in a similar fashion as the Black group as previously described. In this way, the DA estimate for non-Blacks is maintained as the standard for AP3. For non-Black Hispanics under AP2, the non-Black F_{DA}'s were used. Then the AP2 values for Rest were obtained by setting the undercount rates equal to that of the non-Black Hispanics by each of the ten U.S. categories. In summary, use of the F_{DA} adjustments provide a more pronounced differential undercount, for example, between Black male and female in AP2 than in AP1. Substitution is used in all three artificial populations to distribute the population geographically while the F_{DA} factors adjust for age-sex-race differences. If one believes non-Black Hispanic undercount is similar to the Rest group, then AP2 is the relevant population. If one believes they are similar to Blacks in undercount, then AP3 is relevant. A brief numerical summary can be found in Table 1 located in Section 5. The artificial populations exhibit some enumeration overcount for certain categories, for example, Black persons over age 65. Hence, if DA estimates are believed to be correct, the census overenumerated such persons and a negative F_{DA} factor resulted in AP2 and AP3 exhibiting overcounts.

3.2. *Measures of performance*

All census defined geographic areas larger than EDs are unions of EDs. Hence, having defined three artificial populations at the ED level we have three sets of standards with which to compare the performance of the census (enumeration) and the adjustments to be described in the next section. A number of measures of performance were developed and additional ones suggested by others (Citro and Cohen (1985) and Spencer (1986)) were also used. In defining the measures, c represents the enumeration (census), e represents an estimate of the population, s represents the artificial population used as the standard and N denotes the number of areas. The measures consider both estimates of level (total population) as well as proportion of the population.

Measures of performance are:

1. Number of areas where
   $\mathrm{ARE}(c_i) < \mathrm{ARE}(e_i)$
   where
   $\mathrm{ARE}(e_i) = |(e_i - s_i)/s_i|$
   (ARE = absolute relative error)

2. Number of areas where
   $\mathrm{ADP}(c_i) < \mathrm{ADP}(e_i)$
   where
   $\mathrm{ADP}(e_i) = |P_i^e - P_i^s|$
   and
   $P_i^e = e_i / \sum_i^N e_i$ for the ith area
   (ADP = absolute difference in proportions)

3. Number of states erroneously apportioned
   Representation of states in the national legislature is determined by the total population of each state. Using the apportionment formula, this measure provides a count of the number of states affected by adjustment when compared to the standard.

4. $\mathrm{MARE}\ (e) = \frac{1}{N} \sum_i^N \left| \frac{e_i - s_i}{s_i} \right|$
   (MARE = mean ARE)

5. Maximum ARE(e)

6. Median ARE(e)

7. Weighted squared relative error

$$\alpha(e) = \sum_{i}^{N} s_i [(e_i - s_i)/s_i]^2$$

8. Sum of absolute difference of proportions

$$\text{SADP}(e) = \sum_{i}^{N} |P_i^e - P_i^s|$$

9. Proportion of population improved

$$\text{PI}(e) = \sum_{i}^{N} \text{IMPV}_i/\text{M}$$

$$\text{M} = \sum_{i}^{N} s_i,$$

$$\text{IMPV}_i = \begin{cases} s_i & \text{if ADP}(e_i) < \text{ADP}(c_i) \\ 0 & \text{otherwise} \end{cases}$$

10. Weighted squared relative error differences

$$\phi(e) = \sum_{i}^{N} s_i [\{(e_i - s_i)/s_i\} - \{(\sum_{i}^{N} e_i - \sum_{i}^{N} s_i)/\sum_{i}^{N} s_i\}]^2$$

11. Sum of weighted squared ADP

$$\text{IMP1}(e) = \sum_{i}^{N} [\text{ADP}(e_i)]^2/P_i^s$$

4. Adjustment Methods

4.1. *Basic adjustment methods*

The two basic adjustment techniques we have considered are the synthetic and regression techniques. Pertinent references in this area of census adjustment are Purcell and Kish (1979) who provide a general overview of small area estimation techniques, Ericksen and Kadane (1985) who suggest several techniques for census adjustment, Tukey (1981) who proposed conducting census adjustment by use of the blocking concept found in experimental design and Freedman and Navidi (1986) who provide a detailed critique of the models of Ericksen and Kadane (1985). Other references are Cowan and Fay (1984), Bailar (1985) and Tukey (1984), who consider the errors in the 1980 PEP. Other references such as Diffendal, Isaki, and Malec (1982), Isaki, Schultz, Smith, and Diffendal (1987), Isaki, Diffendal, and Schultz (1986), Schultz, Huang, Diffendal, and Isaki (1986), consider the results of applying census undercount adjustment techniques for small areas. Schirm and Preston (1987) assume stochastic models in comparing synthetic state estimators of total population using national level totals by Black/non-Black. They concluded that synthetic estimates perform better than the census with respect to their measures of performance and their standard.

Besides the two basic adjustment techniques, data sources must also be considered. The two components, adjustment techniques and data sources, constitute an adjustment method. Indeed, combinations of techniques and data sources are also possible strategies leading to a plethora of possible adjustment methods. We present the results of such methods in Section 5. All the results that follow assume that all direct estimates (sample based) of the true population are unbiased. That is, all dual system estimates provided by the PES are unbiased.

4.2. *Synthetic estimator*

The idea underlying synthetic estimation is simple. Let $f_{i,\alpha}$ denote the ratio of the true number of category i persons in area α and the census count of the number of such persons in area α. We call $f_{i,\alpha}$ an adjustment factor. Presume that the (i,α)-groups are mutually exclusive and exhaustive in coverage of the population categories and geographic areas of interest. Let $C_{i,\gamma}$ denote the census count of persons in population category i and geographic area γ where area γ is contained in the union of several α areas. Then a synthetic estimator (E) of total population for area γ is given by E_γ,

$$E_\gamma = \sum_{i,\alpha} f_{i,\alpha} \cdot C_{i,\gamma \cap \alpha}, \qquad (1)$$

where the summation is over all categories i

and areas α containing at least a part of area γ ($\gamma \cap \alpha$ denotes intersection of areas α and γ). For example, let the population categories i be Black, non-Black Hispanic and Rest persons (non-Black and non-Hispanic) and the areas α denote place size categories of 0 to 10 000; 10 000 to 100 000 and more than 100 000 within a division of the U.S. Assume that all i categories are crossed by all categories α. If γ is a county in that division containing all three types of population and all three types of areas, E_γ will consist of nine components. The assumption underlying synthetic estimation is that for the areas γ of interest the adjustment factor $f_{i,\alpha\cap\gamma}$ is equal to $f_{i,\alpha}$ for all i's and α's. To the extent that this is not true, E_γ is a potentially biased estimator. The size of the bias has been a topic of some interest. Given an undercount, the reader is reminded that the census is also biased so the issue is then one of comparing the performances of both.

We originally constructed three synthetic estimators labelled syn 1, syn 2 and syn DA. Syn DA utilized adjustment factors defined by age-race-sex (30 factors in total) at the national level only. Syn 2 utilized adjustment factors with the most geographic detail. Adjustment factors were defined within each of nine census geographic divisions (groups of contiguous states) by size and type of place and by three race groups (96 factors in total). In syn 2, no age-sex categories were used to define adjustment factors and occasionally race groups were combined in defining an adjustment factor. Syn 2 uses 96 groups. Syn 1 was a compromise between syn 2 and syn DA in that five geographic categories crossed by 18 age-race-sex categories defined the adjustment factors. Since syn 1 did not perform as well as the other two synthetic estimators, we omit it from further discussion. A complete description of the adjustment factors can be found in Isaki, Diffendal, and Schultz (1987).

4.3. *Regression estimator*

For ease of discussion, we introduce regression estimation in the small area adjustment context assuming modelling of estimated percent net undercount of total population of N areas. Let $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ denote the vector of true net undercount (rate) ($Y_i = (s_i - c_i)/s_i$)) and assume that the regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{O}, \sigma^2\mathbf{I}) \tag{2}$$

holds, where $\mathbf{X}$ is an $N \times p$ matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector. Assume also, that a PES is conducted to measure $\mathbf{Y}$ and let, conditional on $\mathbf{Y}$, $\hat{\mathbf{Y}} \sim N(\mathbf{Y}, \mathbf{D})$ where $\mathbf{D}$ is a diagonal variance-covariance matrix whose diagonal elements d_i are the sampling variances of $\hat{\mathbf{Y}}$. Combining, we have

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \text{ where } \mathbf{u} \sim N(\mathbf{O}, \sigma^2\mathbf{I} + \mathbf{D}). \tag{3}$$

In our work we have assumed that $\mathbf{D}$ is known and estimated σ^2 and $\boldsymbol{\beta}$ using maximum likelihood methods.

Diffendal, Isaki, and Malec (1982) considered using the estimated model in (3) to predict net undercount for smaller areas than states, namely counties. Ericksen and Kadane (1985) using a hierarchical Bayesian framework developed by Lindley and Smith (1972) proposed the estimation of net undercount of states, cities, and balance of states using $\hat{\mathbf{Y}}_{EK}$ where

$$\hat{\mathbf{Y}}_{EK} = [\mathbf{D}^{-1} + \sigma^{-2}\mathbf{I}]^{-1}\,[\mathbf{D}^{-1}\,\hat{\mathbf{Y}} + \sigma^{-2}\mathbf{X}\hat{\boldsymbol{\beta}}] \tag{4}$$

and the use of $\mathbf{X}\hat{\boldsymbol{\beta}}$ to estimate for smaller areas (those areas where PES are not available or where sampling error is too large). Fuller and Harter (1987), using a variance components model approach, obtained similar results. Section 6 contains some results from the modelling of estimated syn 2 adjustment factors, $f_{i,\alpha}$, as well as the modelling of net undercount of states.

5. Synthetic Estimation Results

5.1. *No sampling error*

There are many combinations of possible strategies based on the synthetic and regression techniques. In the remainder of this section we summarize the results previously published concerning statistical synthetic estimation. The reader is referred to Isaki, Diffendal, and Schultz (1986) and Schultz et al. (1986) for details. The results are based on assuming that, separately, artificial populations, AP1, AP2 and AP3 are the true populations. The synthetic estimators investigated are denoted syn 2 and syn DA.

Table 1 below provides a brief summary of net undercount parameters by race at the national level, the adjustment factors for syn DA and the adjustment factors, $f_{i,\alpha}$, for the West South Central division for syn 2 for AP1, AP2 and AP3. Other parameters of interest may be obtained from the authors upon request.

Table 1. Net undercount and adjustment factors for syn DA and syn 2 for AP1, AP2, and AP3[1]

	AP1	AP2	AP3
Net undercount by race at the national level			
1. Total population	.0143	.0162	.0163
2. Black	.0263	.0652	.0652
3. Non-Black Hispanic	.0221	.0151	.0595
4. Rest	.0119	.0089	.0054
Adjustment factors for syn DA			
1. Black, male, 0–14	1.0260	1.0709	1.0709
2. Black, male, 15–29	1.0265	1.0890	1.0890
3. Black, male, 30–44	1.0271	1.2072	1.2072
4. Black, male, 45–64	1.0282	1.1440	1.1440
5. Black, male, 65+	1.0284	.9648	.9648
6. Black, female, 0–14	1.0259	1.0661	1.0661
7. Black, female, 15–29	1.0272	1.0323	1.0323
8. Black, female, 30–44	1.0265	1.0511	1.0511
9. Black, female, 45–64	1.0284	1.0303	1.0303
10. Black, female, 65+	1.0300	.9864	.9864
11. Non-Black Hispanic, male, 0–14	1.0222	1.0037	1.0608
12. Non-Black Hispanic, male, 15–29	1.0230	1.0148	1.0777
13. Non-Black Hispanic, male, 30–44	1.0228	1.0310	1.1749
14. Non-Black Hispanic, male, 45–64	1.0230	1.0213	1.1177
15. Non-Black Hispanic, male, 65+	1.0252	.9972	.9687
16. Non-Black Hispanic, female, 0–14	1.0223	1.0029	1.0571
17. Non-Black Hispanic, female, 15–29	1.0226	.9994	1.0271
18. Non-Black Hispanic, female, 30–44	1.0215	1.0068	1.0417
19. Non-Black Hispanic, female, 45–64	1.0228	1.0084	1.0245
20. Non-Black Hispanic, female, 65+	1.0255	1.0057	.9884
21. Rest, male, 0–14	1.0109	1.0037	.9972
22. Rest, male, 15–29	1.0127	1.0148	1.0091
23. Rest, male, 30–44	1.0123	1.0310	1.0207
24. Rest, male, 45–64	1.0121	1.0213	1.0165
25. Rest, male, 65+	1.0131	.9972	.9982
26. Rest, female, 0–14	1.0109	1.0029	.9965
27. Rest, female, 15–29	1.0123	.9994	.9970
28. Rest, female, 30–44	1.0115	1.0068	1.0042
29. Rest, female, 45–64	1.0121	1.0084	1.0076
30. Rest, female, 65+	1.0138	1.0057	1.0062

Table 1. (cont.) Net undercount and adjustment factors for syn DA and syn 2 for AP1, AP2, and AP3[1]

	AP1	AP2	AP3
Adjustment factors, $f_{i,\alpha}$, for syn 2 for the West South Central Division (Texas, Oklahoma, Arkansas, Louisiana)[1]			
1. Houston and Dallas cities Black	1.0429	1.1126	1.1126
2. Houston and Dallas cities Non-Black Hispanic	1.0403	1.0273	1.1159
3. Houston and Dallas cities Rest	1.0215	1.0170	1.0112
4. Central cities 250 000 + Rest	1.0152	1.0116	1.0077
5. Central cities 50 000–250 000 Rest	1.0153	1.0111	1.0072
6. Areas in 4. and 5. Black	1.0200	1.0506	1.0506
7. Areas in 4. and 5. Non-Black Hispanic	1.0217	1.0138	1.0583
8. In metropolitan area not in central city Rest	1.0185	1.0140	1.0089
9. Cities 10 000–50 000 Rest	1.0140	1.0103	1.0068
10. Rural Rest	1.0206	1.0148	1.0097
11. Areas in 8., 9. and 10. Black or Hispanic	1.0261	1.0405	1.0680

[1] Only a partial listing of syn 2 adjustment factors is presented due to space limitations.

The reader may note that sample based estimates of adjustment factors will eventually be needed for synthetic estimation. We present the results assuming no sampling error because they represent the ideal situation as far as synthetic estimation is concerned. The results provide benchmark standards. Under the conditions given, if synthetic adjustment was found inferior to the census then it would remain inferior in the presence of sampling error. Because the results below were encouraging with respect to synthetic adjustment, examination of the effect of sampling errors on adjustment followed as a logical next step.

Table 2 illustrates the measures of performance of syn 2, syn DA and the census using artificial populations AP1, AP2 and AP3 when total population for states is of interest. The results assume that the $f_{i,\alpha}$ for the statistical synthetic estimators are measured without error. The results in Table 2 revealed that syn 2 and syn DA performed better than the census for all three artificial populations. Furthermore, syn 2 almost always performed better than syn DA.

Table 2. Measures of performance of synthetic estimators compared to the census at the state level (based on 51 states) using artificial populations 1, 2, and 3 for total population

Total population – AP1 Measure No./Description	Syn 2	Syn DA	Census
1 – No. of states where ARE(c_i) < ARE(e_i)	4	7	–
2 – No. of states where ADP(c_i) < ADP(e_i)	12	13	–
3 – Apportionment	2	2	2
4 – MARE	.0042	.0052	.0134
5 – Max ARE	.0147	.0190	.0398
6 – Median ARE	.0028	.0048	.0121
7 – α	4488	8533	52221
8 – SADP	.0031	.0048	.0052
9 – PI	.830	.654	–
10 – ϕ	4488	8211	9735
11 – IMP 1 × 10^{+3}	.0201	.0367	.0447

Total population – AP2 Measure No./Description	Syn 2	Syn DA	Census
1 – No. of states where ARE(c_i) < ARE(e_i)	5	8	–
2 – No. of states where ADP(c_i) < ADP(e_i)	15	14	–
3 – Apportionment	0	2	6
4 – MARE	.0044	.0053	.0147
5 – Max ARE	.0200	.0297	.0771
6 – Median ARE	.0026	.0047	.0113
7 – α	6179	9925	77313
8 – SADP	.0037	.0049	.0067
9 – PI	.703	.694	–
10 – ϕ	6179	9758	17368
11 – IMP 1 × 10^{+3}	.0271	.0429	.0788

Total population – AP3 Measure No./Description	Syn 2	Syn DA	Census
1 – No. of states where ARE(c_i) < ARE(e_i)	7	6	–
2 – No. of states where ADP(c_i) < ADP(e_i)	8	8	–
3 – Apportionment	2	4	8
4 – MARE	.0045	.0047	.0136
5 – Max ARE	.0228	.0300	.0773
6 – Median ARE	.0026	.0032	.0092
7 – α	5866	9344	82339
8 – SADP	.0033	.0047	.0078
9 – PI	.872	.715	–
10 – ϕ	5810	9266	22048
11 – IMP 1 × 10^{+3}	.0255	.0407	.1001

We omitted use of AP1 from further work because AP2 and AP3 were more likely to approximate census undercount by race and because AP1 lacked an age-sex differential that was observed in past censuses.

When considering adjustment for an area smaller than a state, namely a county, we found little guidance to choose between syn DA and syn 2 for both AP2 and AP3, but both are superior to the census. The results are presented in Table 3 below. Syn DA had the smallest MARE and median ARE measures. While syn 2 had a smaller SADP measure, the number of counties in which the census was

Table 3. Measures of performance of synthetic estimators compared to the census at the county level (based on 3137 counties) using artificial populations 2 and 3 for total population

AP2 Measure No./Description	Syn 2	Syn DA	Census
1 – No. of counties where $ARE(c_i) < ARE(e_i)$	1219	1201	–
2 – No. of counties where $ADP(c_i) < ADP(e_i)$	917	870	–
4 – MARE	.0089	.0086	.0128
5 – Max ARE	.2131	.2192	.2236
6 – Median ARE	.0056	.0056	.0076
7 – α	31218	37825	115755
8 – SADP	.0074	.0086	.0115
9 – PI	.736	.703	–
10 – ϕ	31218	37657	55525
11 – IMP 1 $\times 10^{+3}$	.1371	.1656	.2526

AP3 Measure No./Description	Syn 2	Syn DA	Census
1 – No. of counties where $ARE(c_i) < ARE(e_i)$	1325	1266	–
2 – No. of counties where $ADP(c_i) < ADP(e_i)$	723	707	–
4 – MARE	.0081	.0074	.0111
5 – Max ARE	.2946	.2757	.3067
6 – Median ARE	.0044	.0039	.0055
7 – α	34688	41508	134577
8 – SADP	.0071	.0084	.0131
9 – PI	.783	.747	–
10 – ϕ	34633	41430	74347
11 – IMP 1 $\times 10^{+3}$	.1519	.1821	.3366

superior to syn 2 was not much different from syn DA. Likewise the PI measures were similar. The universe of counties were divided by population into three size groups 0 to 10 000; between 10 000 and 50 000; and those exceeding 50 000 with about 25%, 50% and 25% of the counties, respectively. Each of the three groups were looked at separately. This analysis indicated that syn DA fared well for the smaller population size, syn 2 fared well for the larger population size and for the middle group the results were mixed.

While syn 2 was better for states and syn DA for counties, the observation that syn 2 was also superior for large counties suggests that there may not be a single synthetic estimator satisfactory for all areas. Instead, we may need to apply separate synthetic estimators over portions of the universe of all areas.

A second consideration is that of sampling error which is covered in the next section. Since in practice the adjustment factors need to be estimated, the sampling error of the synthetic estimators warrant consideration.

5.2. *Sampling error*

In order to examine the sampling error effect, we devised a simple sample design using EDs as the sampling unit even though it is likely that a smaller unit such as a census block is likely to be used in a PES for 1990. The ED was the smallest geographical unit on our data file. For each ED, counts by race-age-sex were available for the 1980 census and our artificial population variables AP2 and AP3.

The sample design was constructed to support estimation of the 96 adjustment factors of

syn 2. In this respect, the universe of EDs was stratified along adjustment factor definitions (Huang (1987)). The sample number of EDs was set at 1 440. This number was determined by assuming that an ED contained on average seven blocks and hence approximates a 10 000 block sample design that had been suggested as a rough sample size for a PES in 1990. Sample sizes of EDs were allocated proportionally to the population of the sampling stratum. The sampling weights were approximately 200. The EDs were assigned to sampling strata on the basis of geography and 1980 census percent minority category. Because of this, some sample estimated adjustment factors within census division are correlated but they are never correlated between divisions. Due to the limitations of the computer, 90 replicates were selected, each containing 1 440 EDs. Each replicate represents a potential sample realization; the replicates were obtained via equal probability systematic sampling. From each replicate a set of 96 adjustment factors for syn 2 and a set of 30 adjustment factors for syn DA were computed. These replicates were used to compute covariance matrices for each set of adjustment factors.

We chose one of the 90 replicates and computed the summary measures on total population for AP2 and AP3 for both syn 2, syn DA and the census when states and counties are of interest. The covariance matrix will be used to compute mean square errors of the synthetic estimators as well as to study regression methods in small area estimation. Because syn 2 requires more parameters to be estimated, its summary measures are expected to be affected to a larger degree than those for syn DA. When viewed in the 1990 context some inefficiency in the sample design used, e.g., EDs versus blocks, is balanced by the fact that we used the current census data which for 1990 will not be available. The net effect of this balancing of conditions is not known. The sampling error effect is due to the variability of a poststratified estimator of the numerator of the adjustment factor where the value of each sample selected ED is assumed to be measured without error. The sampling error effect is not a measure of the variability of a dual system estimator.

Using a single replicate, randomly selected from the group of 90 and applying the resulting state and county adjusted figures to the measures of performance we computed Tables 4 and 5. Only a single replicate was used because of cost. Adjustment results for each replicate required adjusting census counts at the ED level and tabulating to the area of interest. Basically, the results for both state and counties favor syn DA over syn 2 and the census. As can be seen in Table 4, the sampling variability of the estimated adjustment factors has caused syn 2's performance to diminish relative to syn DA although both remain superior to the census. One way to reduce the variability of the adjustment factors is the use of regression models in the manner of Ericksen and Kadane (1985) and Fuller and Harter (1987). We illustrate the results of application of their methods in the next section.

6. Regression Results

6.1. Smoothing estimated syn 2 adjustment factors

To improve the directly estimated adjustment factors of syn 2, a regression model along the lines of equation (3) was applied to the adjustment factors, $f_{i,\alpha}$, from a single replicate. The 96 sample estimated syn 2 adjustment factors were modelled using equation (3) with the only available explanatory variables – namely, classification variables such as geographic division, size of place, and race. The resulting regression models for AP2 and AP3, respectively, are

$$\hat{y}_i^{AP2} = 1.019 + 0.04\, X_{Bi} - 0.007\, X_{Ri} - 0.006\, X_{13i}$$
$$\hat{\sigma}^2_{AP2} = .939 \times 10^{-5}, \qquad (5)$$

$$\hat{y}_i^{AP3} = 1.008 + 0.052\, X_{NWi} - 0.004\, X_{12i}$$
$$\hat{\sigma}^2_{AP3} = .429 \times 10^{-5}, \qquad (6)$$
$$i = 1, \ldots, 96$$

where

$$X_{13i} = X_{1i} + X_{2i} + X_{5i} + X_{6i} + X_{8i},$$
$$X_{12i} = X_{1i} + X_{2i} + X_{4i} + X_{5i} + X_{6i} + X_{8i},$$

y_i is the adjustment factor,
X_{ji} is the indicator variable for geographic division j, $j = 1, \ldots, 9$,
X_{Bi} is the population proportion of Black persons in the adjustment category i,
X_{Ri} is the population proportion of Rest persons (non-Black, non-Hispanic) in the adjustment category i,
X_{NWi} is the population proportion of non-White (Black and Hispanic) persons in category i.

Race and geographic division were the explanatory variables selected in the regression of syn 2 adjustment factors. The difference between the artificial populations are reflected in the difference of selection of explanatory variables in the regression of adjustment factors. AP2 construction assumed Hispanics were similar to the Rest of the population; Black and Rest were found to be significant in predicting adjustment factors. AP3 assumed that Hispanics and Blacks were similar; the non-White variable was found to be a better predictor than Black alone in predicting the adjustment factor. X_{13} is an area indicator variable for most of the east coast, inland and mountain states; X_{12} includes X_{13} and geographic division 4 which includes four southern states. The predicted adjustment factors for each category i for AP2 and AP3 are estimated by using equation (4) with estimated σ^2 for AP2 and AP3 respectively. The adjusted census counts are then evaluated at state and county levels using the measures of performance previously defined. (See Tables 4 and 5). The syn 2 estimator with adjustment factors as a result of (4) (termed Smoothed Factors in Tables 4 and 5) showed some improvement over the syn 2 estimator without modelled adjustment factors for almost all the measures at both state and county level.

6.2. Alternative-smoothing of syn 2 estimated undercount at the state level

An alternative to smoothing the adjustment factors via regression and then applying the smoothed factors to the census counts is as follows. First, apply the directly estimated factors to the census counts as in syn 2 and produce adjusted counts for states. Second, estimates of the net undercount for states can then be modelled using explanatory variables produced from census tabulations also at the state level. The smoothed state net undercount estimates are then converted to state level population estimates. One reason for looking at this alternative approach was that there were very few explanatory variables available for use in forming a model to smooth the adjustment factors. Very few tabulations at the adjustment factor level existed whereas there were many more explanatory variables tabulated to the state level.

A model was constructed for each artificial population based on assumptions in (2) and (3). The models selected were

$$\hat{Y}_i^{AP2} = -.709 + .224\, Z_{Ai} + .096\, Z_{Ri} \qquad (7)$$
$$\hat{\sigma}^2_{AP2} = .083$$
$$\hat{Y}_i^{AP3} = -.257 + .069\, Z_{Mi} + .094\, Z_{Ai} \qquad (8)$$
$$\hat{\sigma}^2_{AP3} = .003,$$

where

Y_i is the percent net undercount at the state level,
Z_{Ai} is the percent item imputations,
Z_{Ri} is the percent minority renters, and
Z_{Mi} is the percent minority.

Table 4. Measures of performance of synthetic estimators compared to the census at the state level (51) for total population using artificial populations 2 and 3 for a single replicate

AP2

Measure No./Description	Syn 2	Smoothed factors	Smoothed state	Smoothed EK Bayes	Syn DA	Census
1 – No. of states where ARE (c_i) < ARE(e_i)	6	5	5	4	8	–
2 – No. of states where ADP(c_i) < ADP(e_i)	20	14	11	12	13	–
3 – Apportionment	2	2	2	2	2	6
4 – MARE	.0060	.0054	.0039	.0040	.0053	.0147
5 – Max ARE	.0218	.0350	.0131	.0125	.0288	.0771
6 – Median ARE	.0039	.0032	.0025	.0028	.0048	.0113
7 – α	12189	9333	8160	6630	9282	77313
8 – SADP	.0056	.0046	.0045	.0043	.0048	.0067
9 – PI	.481	.687	.646	.677	.757	–
10 – φ	11985	8823	8007	6477	9282	17368
11 – IMP 1 × 10^{+3}	.0525	.0388	.0351	.0284	.0408	.0788

AP3

Measure No./Description	Syn 2	Smoothed factors	Smoothed state	Smoothed EK Bayes	Syn DA	Census
1 – No. of states where ARE (c_i) < ARE(e_i)	8	7	7	7	8	–
2 – No. of states where ADP(c_i) < ADP(e_i)	17	11	12	12	9	–
3 – Apportionment	4	4	4	4	4	8
4 – MARE	.0060	.0048	.0041	.0041	.0049	.0136
5 – Max ARE	.0362	.0355	.0186	.0186	.0290	.0773
6 – Median ARE	.0038	.0021	.0023	.0023	.0033	.0092
7 – α	19227	10608	7650	7650	9180	82339
8 – SADP	.0068	.0046	.0042	.0042	.0046	.0078
9 – PI	.635	.696	.639	.639	.703	–
10 – φ	18968	9843	7650	7650	9129	22048
11 – IMP 1 × 10^{+3}	.0822	.0433	.0335	.0335	.0401	.1000

Table 4 gives the results of the measures of performance defined earlier at the state level. In the table, the census results can be compared with both the level results derived from equations (7) and (8) above (termed smoothed state) and the predictions based on the weighted average of the direct estimate with the modelled results (termed smoothed EK Bayes in equation (4)) for both artificial populations. Comparisons can also be made to syn 2, to the smoothed factors model and to syn DA when estimating for states.

To investigate the synthetic regression assumption, equations (7) and (8) were applied to county level data and the measures of performance computed to see how well the state models did in predicting county results. As can be seen from Table 5 the use of the synthetic assumption to predict county estimates did better than the census when compared to the artificial populations. Table 5 also indicates that the smoothed state model combined with the synthetic regression assumption is superior to the smoothed factors model. The interested reader is referred to Isaki and Schultz (1987b) and Schultz, Isaki, and Diffendal (1987) for further results concerning regression models and their performance.

Table 5. Measures of performance of synthetic estimators compared to the census at the county level (3137) for total population using artificial populations 2 and 3 for a single replicate

AP2

Measure No./Description	Syn 2	Smoothed factors	Smoothed state	Syn DA	Census
1 – No. of counties where $ARE(c_i) < ARE(e_i)$	1104	1194	1048	1254	–
2 – No. of counties where $ADP(c_i) < ADP(e_i)$	999	947	864	862	–
4 – MARE	.0092	.0087	.0081	.0087	.0128
5 – Max ARE	.2200	.2179	.2072	.2192	.2236
6 – Median ARE	.0052	.0052	.0047	.0058	.0076
7 – α	44859	37958	33566	36703	115755
8 – SADP	.0093	.0086	.0078	.0085	.0115
9 – PI	.625	.710	.698	.702	–
10 – ϕ	44515	37330	33566	36703	55525
11 – IMP $1 \times 10^{+3}$	.1953	.1650	.1471	.1617	.2526

AP3

Measure No./Description	Syn 2	Smoothed factors	Smoothed state	Syn DA	Census
1 – No. of counties where $ARE(c_i) < ARE(e_i)$	1122	1121	1154	1358	–
2 – No. of counties where $ADP(c_i) < ADP(e_i)$	821	721	753	702	–
4 – MARE	.0081	.0075	.0075	.0077	.0111
5 – Max ARE	.3007	.2998	.2672	.2720	.3067
6 – Median ARE	.0042	.0040	.0035	.0044	.0055
7 – α	61485	44545	37330	41095	134577
8 – SADP	.0098	.0087	.0078	.0084	.0131
9 – PI	.680	.762	.736	.743	–
10 – ϕ	61172	43918	37017	41045	74347
11 – IMP $1 \times 10^{+3}$	.2676	.1934	.1630	.1798	.3366

More work is needed to determine the suitability of regression models versus synthetic estimators for different publication levels. We have partial evidence that, within some states and for some measures, the census performs as well as the synthetic estimator at the ED level. The question remains as to adjustment of intermediate sized areas. A parallel issue is the smoothing of adjustment factor estimates. We did not utilize in the regression modelling of adjustment factors some of the possibly useful explanatory variables. Smoothing of adjustment factors has the advantage of requiring less interaction with census data over the smoothed state approach. For example, having adjusted county totals via the smoothed state approach requires an additional step to distribute the estimate over subcounty areas.

We intend to focus future research on a single state, New Jersey, with respect to subcounty adjustment down to the block level. We must also devise adjustment procedures to carry our direct estimates down to small areas based on the results to date. This task begins with determining the best combined estimate of parameters using demographic analysis, PES, and the census. All of these activities are now a part of the evaluation program of the 1990 census.

7. Conclusion

The paper began with a brief overview of census adjustment, i.e., the goals and the available undercount measurement sources. Unable to assess potential adjustment methods, we resorted to constructing artificial populations and using measures of performance that seemed reasonable. Interpretation and application of the measures as indicating improvement in adjustment over the census is left to the reader. Clearly, determining the superiority of one method over another is subjective when the measures of performance differ. Our aim was to construct plausible adjustment methods and present the results.

A clear limitation of this paper is its dependency on the artificial populations since the undercount in 1990 may differ substantially. However, the artificial populations are reasonable proxies for the undercount to evaluate different small area estimates. Another is the simple sampling design used to approximate sampling error effects in adjustment. Finally, direct estimates of total population are assumed to be unbiased (Freedman and Navidi (1986)). The strength of this paper is that it provides numerical illustrations of the performance of some adjustment methods compared to the census.

At the risk of excessively generalizing the results of adjustment on the basis of the artificial populations, we have the following observations.

i. Synthetic estimation provides better measures of performance than the census at the state and county level. The result holds even in the presence of sampling error.
ii. Regression models reduce the effects of sampling error and improve adjustment results.
iii. As the adjustment results hold for counties, especially the larger counties (total population exceeding 50 000), it is speculated that administrative areas of comparable size such as places and central cities of metropolitan areas will also favor adjustment.
iv. An adjustment for undercount does not improve every county's or state's population figures. However, more areas will be improved than will be degraded through an adjustment.

8. References

8.1. References cited in the text

Bailar, B. A. (1985): Comments on Estimating the Population in a Census Year by E. P. Ericksen and J. B. Kadane, Journal of the American Statistical Association, 80, pp. 109–114.

Citro, C. F. and Cohen, M. L. (1985): The Bicentennial Census – New Directions for Methodology in 1990. National Academy Press.

Cowan, C. D. and Bettin, P. J. (1982): Estimates and Missing Data Problems in the Postenumeration Program. Technical report, U.S. Bureau of the Census, Washington, D.C.

Cowan, C. D. and Fay, R. E. (1984): Estimates of Undercount in the 1980 Census. American Statistical Association, Proceedings of the Survey Research Methods Section, pp. 560–565.

Cowan, C. D. and Malec, D. J. (1986): Capture-Recapture Models When Both Sources Have Clustered Observations. Journal of the American Statistical Association, 81, pp. 347–353.

Diffendal, G. J., Isaki, C. T., and Malec, D. J. (1982): Examples of Some Adjustment Methodologies Applied to the 1980 Census. Technical report, U.S. Bureau of the Census, Washington, D.C.

Ericksen, E. P. and Kadane, J. B. (1985): Estimating the Population in a Census Year – 1980 and Beyond. Journal of the Amer-

ican Statistical Association, 80, pp. 98–109.

Freedman, D. A. and Navidi, W. C. (1986): Model for Adjusting the Census. Statistical Science, 1, pp. 3–11.

Fuller, W. A. and Harter, R. (1987): The Multivariate Components of Variance Model for Small Area Estimation. In Small Area Statistics – An International Symposium, John Wiley and Sons, New York.

Huang, E. T. (1987): Survey Based Estimates of Census Adjustment Factors and Its Variances and Covariances – A Monte Carlo Study. Statistical Research Division Report Series, CENSUS/SRD/RR-87/08, U.S. Bureau of the Census, Washington, D.C.

Isaki, C. T. (1986): Bias of the Dual System Estimator and Some Alternatives. Communications in Statistics – Theory and Methods, 15 (5), pp. 1435–1450.

Isaki, C. T., Diffendal, G. J., and Schultz, L. K. (1986): Statistical Synthetic Estimates of Undercount for Small Areas. Proceedings of the U.S. Census Bureau's Second Annual Research Conference, pp. 557–569.

Isaki, C. T., Diffendal, G. J. and Schultz, L. K. (1987): Report on Statistical Synthetic Estimation for Small Areas. Statistical Research Division Report Series, CENSUS/SRD/RR-87/02, U.S. Bureau of the Census, Washington, D.C.

Isaki, C. T. and Schultz, L. K. (1987a): The Effects of Correlation and Matching Error in Dual System Estimation. Communications in Statistics – Theory and Methods, 16 (8), pp. 2405–2427.

Isaki, C. T. and Schultz, L. K. (1987b): Report on the Effects of the Violations of Assumptions on Regression Estimation of Census Coverage Error. Statistical Research Division Report Series, CENSUS/SRD/RR-87/04, U.S. Bureau of the Census, Washington, D.C.

Isaki, C. T., Schultz, L. K., Smith, P. J., and Diffendal, G. J. (1987): Small Area Estimation Research for Census Undercount – Progress Report. In Small Area Statistics – An International Symposium, John Wiley and Sons, New York.

Jabine, T. and Bershad, M. (1970): Some Comments on the Chandrasekaran-Deming Technique for the Measurement of Population Change. Published by the Office of United States Economic Coordinator for CENTO Affairs. Proceedings of the CENTO Symposium on Demographic Statistics, pp. 189–206.

Jaro, M. (1985): Current Record Linkage Research. Paper presented at the April 1985 Meeting of the Census Advisory Committee, Washington, D.C.

Lindley, D. V. and Smith, A. F. M. (1972): Bayes Estimates for the Linear Model. Journal of the Royal Statistical Society, Ser. B, 34, pp. 1–19.

Purcell, N. J. and Kish, L. (1979): Estimation for Small Domains. Biometrics, 35, pp. 365–384.

Schirm, A. L. and Preston, S. H. (1987): Census Undercount Adjustment and the Quality of Geographic Distributions. Journal of the American Statistical Association, 82, pp. 965–978.

Schultz, L. K., Huang, E. T., Diffendal, G. J., and Isaki, C. T. (1986): Some Effects of Statistical Synthetic Estimation on Census Undercount of Small Areas. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 321–325.

Schultz, L. K., Isaki, C. T., and Diffendal, G. J. (1987): Report on Using Regression Models for Small Area Adjustment. Statistical Research Division Report Series, CENSUS/SRD/RR-87/01, U.S. Bureau of the Census, Washington, D.C.

Seber, G. A. F. (1982): The Estimation of

Animal Abundance and Related Parameters. MacMillan, New York.

Spencer, B. (1986): Conceptual Issues in Measuring Improvement in Population Estimates. Proceedings of the U.S. Census Bureau's Second Annual Research Conference, pp. 393–407.

Tukey, J. W. (1981): Discussion of "Issues in Adjusting the 1980 Census Undercount", by Barbara Bailar and Nathan Keyfitz. Paper presented at the Annual Meeting of the American Statistical Association, Detroit, MI.

Tukey, J. W. (1984): Points to be Made. Presented to the Committee on National Statistics, Panel on Decennial Census Methodology, May 11, 1984.

Wolter, K. M. (1986): Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, pp. 338–346.

8.2. References not cited in the text

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988): An Error Components Model for Prediction of County Crop Areas Using Survey and Satellite. Journal of the American Statistical Association, 83, pp. 28–36.

Citro, C. F. and Pratt, J. W. (1986): The USA's Bicentennial Census: New Directions for Methodology in 1990. Journal of Official Statistics, Vol. 2, pp. 359–381.

Coale, A. J. and Rives, N. W., Jr. (1973): A Statistical Reconstruction of the Black Population of the United States, 1880–1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility. Population Index 39, (1), pp. 3–36.

Coale, A. J. and Zelnik, M. (1963): New Estimates of Fertility and Population in the United States: A Study of Annual White Births from 1855 to 1960 and of Completeness of Enumeration in the Censuses from 1880 to 1960. Princeton University Press, Princeton, N.J.

Isaki, C. T. and Schultz, L. K. (1987c): Report on Demographic Analysis Synthetic Estimation for Small Areas. Statistical Research Division Report Series, CENSUS/SRD/RR-87/03, U.S. Bureau of the Census, Washington, D.C.

Spencer, B. (1980): Benefit-Cost Analysis of Data Used to Allocate Funds. Lecture Notes in Statistics 3. Springer-Verlag, New York.

Spencer, B. (1980): Implications of Equity and Accuracy for Undercount Adjustment: A Decision-Theoretic Approach. Proceedings of the 1980 Conference on Census Undercount, U.S. Department of Commerce, Bureau of the Census, Washington, D.C., pp. 204–216.

Received February 1987
Revised January 1988

Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 113–124

On Autoregressive Model Identification

Ette Harrison Etuk[1]

Abstract: Since Cleveland (1972) introduced the inverse autocorrelation function, it has been recognized as a competitor to the partial autocorrelation function as a time series model identification tool. By using simulated and real data, we have demonstrated that neither of these is consistently more powerful than the other for identification of autoregressive (AR) models. However when the underlying AR process is of full order, the partial autocorrelation function invariably is the superior. But when a subset order AR model generates the data, the inverse autocorrelation function is generally more informative. On the whole the partial autocorrelation function exhibits better performance. For instance, in two of the three cases of real series used it clearly outperforms the inverse autocorrelation function.

Key words: Autoregressive model identification: partial autocorrelation function; inverse autocorrelation function.

1. Introduction

Since Yule (1927) introduced autoregressive modelling, it has played a significant role in the analysis of data recorded sequentially in time or space. The basic component of all time series models is the white noise process defined as a sequence $\{\varepsilon_t\}$ of uncorrelated zero mean and constant variance random variables. A mean corrected time series $\{X_t\}$ is said to be an autoregressive process of order p (designated AR(p)) if it is a stationary solution of the following difference equation.

$$X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} = \varepsilon_t \,, \qquad (1.1)$$

where the α_i 's are constants. For stationarity, the characteristic polynomial $\sum_{i=0}^{p} \alpha_i z^i$, $\alpha_0 = 1$, must have zeros outside the unit circle.

Any stationary time series can be expressed as an infinite-order autoregression. Autoregressive modelling therefore involves the approximation of the autoregression by a finite order one of the form (1.1). Thus the order, in essence, constitutes an additional parameter to be estimated. This order determination represents a major obstacle since underfitting increases the residual variance while overparametrization decreases the reliability of the model. Model identification involves more than order determination; the relative contribution of the parameters to the model structure should also be estimated.

A diagnostic aid for autoregression order determination is the partial autocorrelation function (PACF), advocated for this purpose by Box and Jenkins (1976). It is known to cut off at lag p for (1.1). Hence its estimate

[1] Department of Mathematics and Statistics, University of Calabar, Calabar, Nigeria.

indicates a possibility of an AR(p) component of the underlying model, if it fails to be significant after lag p. A similar tool, the inverse autocorrelation function (IACF), has been introduced by Cleveland (1972).

The question of which is better in AR model identification has engaged the attention of many researchers since the introduction of the IACF. Cleveland (1972), Chatfield (1979), Hipel, McLeod, and Lennox (1977), and Oyetunji (1985) to mention a few, believe that the IACF is the better. However, McLeod, Hipel, and Lennox (1977) observe that in certain time series applications, the PACF is comparatively better in specifying the model. Abraham and Ledolter (1984) have demonstrated that the PACF is more powerful in identifying purely AR processes.

It is the objective of this work to further investigate the relative merits of the two methods. Like Abraham and Ledolter (1984), we simulate AR models and observe the frequency with which the functions detect certain features of the model. We observe the effect of the nature of roots of the characteristic equation of the underlying model, the sample size, and the distance of the model from the boundary of stationarity on their comparative performance. Another factor of variation of interest is whether the model is full order or subset order. Inspired by the work of Bhansali (1983), we use the intuitively appealing autoregressive estimates of the IACF, having generated AR models. Since the estimation of the functions, especially that of the PACF, is inextricably tied to that of the model, the nature of their estimates depends on the mode of model estimation. Our Monte Carlo study uses the Yule-Walker approach to autoregression estimation. However in Section 6 we use also Burg's estimates of the IACF and PACF, to illustrate our results on some real series.

2. Partial Autocorrelation Function (PACF)

Suppose we write (1.1) more specifically as

$$X_t + \alpha_{p1}X_{t-1} + \alpha_{p2}X_{t-2} + \ldots + \alpha_{pp}X_{t-p} = \varepsilon_t \quad (2.1)$$

where α_{ij} is the jth coefficient of an AR(i) model. The last coefficient α_{pp} is called the partial autocorrelation of lag p.

The sequence $\{\alpha_{pp}\}$, regarded as a function of p, is the PACF of $\{X_t\}$.

Let $\mu = E(X_t)$, $\gamma_k = E[(X_t-\mu)(X_{t-k}-\mu)] = \gamma_{-k}$, and $\varrho_k = \frac{\gamma_k}{\gamma_0}$, $k = 0, \pm 1, \pm 2, \ldots$ be the mean, autocovariance of lag k, and autocorrelation of lag k, respectively, of $\{X_t\}$. They can be estimated by

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_t, \; C_k = \frac{1}{N}\sum_{t=k+1}^{N} (X_t-\bar{X})(X_{t-k}-\bar{X}),$$

$$\text{and } r_k = \frac{C_k}{C_0}, \; k = 0, 1, 2, 3, \ldots$$

respectively, from a realization, $X_1, X_2, \ldots, X_N$, of $\{X_t\}$.

Contemporary techniques of AR modelling use automatic order determination criteria whose minimum within a specified order range gives the optimum order.
Criteria like

$$\text{FPE}(p) = \hat{\sigma}_p^2\left(1 + \frac{p}{N}\right), p = 0,1,2,\ldots,$$
(Akaike (1969))

$$\text{FPE}\alpha(p) = \left(1 + \frac{\alpha p}{N}\right)\left(1 - \frac{p}{N}\right)^{-1}\hat{\sigma}_p^2,$$

$$\alpha > 0, p = 0,1,2,\ldots,$$
(Bhansali and Downham (1977))

$$\text{AIC}(p) = N\ln\tilde{\sigma}_p^2 + 2p, p = 0,1,2,\ldots,$$
(Akaike (1977))

$$\text{BIC}(p) = N\ln\tilde{\sigma}_p^2 - (N-p)\ln\left(1-\frac{p}{N}\right) + p\ln N$$

$$+ pln\{p^{-1}\left(\frac{C_o}{\tilde{\sigma}_p^2} - 1\right)\}, p = 0,1,2,\ldots,$$
(Akaike (1977))

$$\mathrm{SIC}(p) = Nln\tilde{\sigma}_p^2 + plnN, p = 0,1,2,\ldots,$$
(Schwarz (1978))

$$\mathrm{CAT}_2(p) = \begin{cases} \frac{1}{N}\left(\sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2}\right) - \frac{1}{\hat{\sigma}_p^2}, p = 1,2,\ldots \\ -\left(1 + \frac{1}{N}\right), p = 0, \end{cases}$$
(Parzen (1977))

$$\mathrm{CAT}_3(p) = \frac{1}{N}\left(\sum_{j=o}^{p} \frac{1}{\hat{\sigma}_j^2}\right) - \frac{1}{\hat{\sigma}_p^2},$$
$p = 0,1,2,\ldots,$ (Tong (1977))

$$S_N(p) = (N + 2p)\,\hat{\sigma}_p^2, p = 0,1,2,\ldots,$$
(Shibata (1980))

$$\phi(p) = ln\tilde{\sigma}_p^2 + N^{-1}2pc\,lnln\,N, c>1, p = 0,1,2,\ldots$$
(Hannan and Quinn (1979))

are to mention a few, where $\hat{\sigma}_p^2$ and $\tilde{\sigma}_p^2$ are the least squares and maximum likelihood estimates of the residual variance, respectively.

An approximate least squares estimate (Yule-Walker estimate) of (2.1) can be obtained by the use of the recursive formula:

$$\phi_{k+1} = \hat{\alpha}_{k+1,k+1} = \frac{r_{k+1} - \sum_{j=1}^{k} \hat{\alpha}_{kj}\, r_{k+1-j}}{1 - \sum_{j=1}^{k} \hat{\alpha}_{k,j}\, r_j},$$

$$\hat{\alpha}_{k+1,j} = \hat{\alpha}_{k,j} - \phi_{k+1}\,\hat{\alpha}_{k,k-j+1}, j = 1, 2,\ldots, k.$$

Another method of autoregression estimation is the maximum entropy method proposed by Burg and formalized by Andersen (1974).

Under the hypothesis of an AR (p) model, it is well known that the estimated partial autocorrelations of order $p+1$ and higher are approximately independently distributed with $E(\hat{\alpha}_{kk}) = 0$, and

$$\mathrm{var}(\hat{\alpha}_{kk}) \cong \frac{1}{N}, k \geqslant p + 1 \qquad (2.2)$$

as $N \to \infty$.

3. Inverse Autocorrelation Function (IACF)

For a stationary stochastic process $\{X_t\}$ with spectral density function $f(\omega)$, Cleveland (1972) has defined the inverse autocovariance of lag k as

$$\gamma i_k = \int_{-\pi}^{\pi} e^{i\omega k} fi(\omega)\, d\omega = \gamma i_{-k}, k = 0, 1, 2,\ldots$$

where $fi(\omega) = \frac{1}{f(\omega)}$. The inverse autocorrelation function is defined as

$$\varrho i_k = \frac{\gamma i_k}{\gamma i_o}, k = 0, 1, 2, \ldots .$$

An equivalent but time-domain definition of γi_k is given by Chatfield (1979).

Using the duality between AR and moving average processes, it has been shown (Cleveland (1972); Chatfield (1979)) that for the AR (p) process (1.1), ϱi_k is given by

$$\varrho i_k = \begin{cases} \left(\alpha_k + \sum_{j=1}^{p-k} \alpha_j\,\alpha_{j+k}\right) / \left(1 + \sum_{j=1}^{p} \alpha_j^2\right) \\ k = \pm 1,\ldots, \pm p \\ 0, |k| > p \end{cases} \qquad (3.1)$$

and

$$\gamma i_o = \left(1 + \sum_{j=1}^{p} \alpha_j^2\right) \sigma_\varepsilon^2, \qquad (3.2)$$

where

$$\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon_t).$$

Cleveland (1972) has suggested two methods of estimating ϱi_k stemming from two methods of spectrum estimation, viz., the autoregressive and window methods. The autoregressive method, which is quite popular, consists of approximating the process by an AR model of sufficiently high order p for a good fit, estimating the parameters of the model and using the estimates in equations (3.1) and (3.2) to obtain an estimate of IACF.

The subjectivity introduced in choosing p poses problems. Hipel, McLeod, and Lennox (1977) suggest trying four values of p between 10 and 40 (where $p < N/4$) and choosing the value of p which gives the most representative graph of the resultant inverse autocorrelation estimate ri_k against lag k. Chatfield (1979) advises against the use of automatic criteria like the AIC, BIC, etc. since for determining p optimal parametric parsimony is not of interest. He however warns that p must not be so large as to make the variance of the estimates too high; the choice of p must be such that for high lags, the estimates of the inverse autocorrelations approach zero. He suggests some form of trial and error until the foregoing criteria are met. Hosking (1980) suggests that p should vary with the sample size N. In their work Abraham and Ledolter (1984) use the values of 5 and 10 for p since their simulated models had lower orders and report the results for $p = 10$ after observing the results for the two values to be similar.

The window method involves smoothing the periodogram. A good exposition on this approach is given by Priestley (1981). Other methods of estimation of IACF include those proposed by McClave (1978) and Chatfield (1979).

Hosking (1980) has shown that for a stationary time series, $N^{1/2}(ri_k - \varrho i_k)$ is normal with mean zero and covariances given by

$$\text{cov}(ri_k, ri_{k+t}) \simeq \frac{1}{N}\sum_{v=-\infty}^{\infty}\{\varrho i_v \varrho i_{v-t} + \varrho i_{v-k-t}\varrho i_{v+k} - 2\varrho i_k \varrho i_{v-k-t} - 2\varrho i_{k+t}\varrho i_v \varrho i_{v-k}\} \text{ as } N \to \infty.$$

Thus for an AR (p) process, asymptotically,

$$\text{var}(ri_k) \simeq \frac{1}{N}\{1+2\sum_{v=1}^{p}\varrho i_v^2\}, k > p. \quad (3.3)$$

4. Simulated Data

Eight AR(2) series with (α_1, α_2) equal to (–1.68, 0.70), (0.00,–0.78), (–0.66, 0.10), (0.00,–0.15), (–1.08, 0.77), (0.00, 0.89), (–0.46, .0.08) and (0.00, 0.10) are simulated independently twenty times each. In the sequel, we shall refer to them as series I trough VIII, respectively. Table 1 shows which of them have characteristic equations with real or complex roots and those close to or far from the boundary of stationarity.

Table 1. Categories of simulated series

	Real roots	Complex roots
Close	I, II	V, VI
Far	III, IV	VII, VIII

The motivation for this choice of models is the need to cover the parameter surface. We also use sample sizes of 50, 150, and 250 for each series.

The white noise process for each simulation is a sequence of pseudorandom numbers generated using the RAN function of the FORTRAN 77 language. The sequence is made approximately standard normal.

5. Results

We, as did Abraham and Ledolter (1984), use the critical region $\pm 2\ N^{-1/2}$ for assessing the performance of both criteria. We also use $p=5$ and $p=10$ and compare the results.

Like that of Abraham and Ledolter (1984) our observation is that the PACF detects the significance of the lag two coefficient more than the IACF for all the models. This is especially true for full-order models, those with characteristic equations having real roots and those close to the boundary of stationarity. Both the performance of each criterion and the relative performance of the IACF increase with sample size. IACF also performs better for $p = 5$ than for $p = 10$. We have observed also that both functions are good at detecting zero third-order components.

To examine their relative power in order determination, we note the frequency with which each has a significant second-lag value and a non-significant third-lag value. Coefficients at higher lags were observed to be non-significant. Table 2 shows our results. Clearly, the PACF excels for all the models and for both lags. The larger the sample size, the better the comparative performance of the IACF.

Comparing their potential for correctly detecting α_1, we find that while the IACF excels for subset-order models, the PACF excels for full-order models. Our result for the subset-order models supports the claims of Cleveland (1972) and Chatfield (1979), to mention a few. The details of the results are in Table 3. The IACF is more powerful especially when the underlying subset-order

Table 2. Frequencies of the events $\{(\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{\alpha}_{33}): |\hat{\alpha}_{22}| > 2N^{-1/2}, |\hat{\alpha}_{33}| \leq 2N^{-1/2}\}$ *and* $\{ri_1, ri_2, ri_3\} : |ri_2| > 2N^{-1/2}, |ri_3| \leq 2N^{-1/2}\}$

		$N=50$		$N=150$		$N=250$		Total	
Series I	IACF	0	0*	1	1*	5	5*	6	6*
	PACF	12		13		10		5	
Series II	IACF	18	20*	18	20*	15	20*	51	60*
	PACF	19		20		18		57	
Series III	IACF	0	0*	0	0*	4	5*	4	5*
	PACF	2		2		11		15	
Series IV	IACF	3	0*	4	1*	10	11*	17	12*
	PACF	3		3		11		17	
Series V	IACF	10	5*	20	20*	20	19*	50	44*
	PACF	19		19		20		58	
Series VI	IACF	20	20*	16	20*	19	19*	55	59*
	PACF	19		20		20		59	
Series VII	IACF	2	1*	5	4*	5	4*	12	9*
	PACF	3		7		5		15	
Series VIII	IACF	0	0*	7	8*	7	5*	14	13*
	PACF	1		6		7		14	
Total		53	46*	71	74*	85	88*		
		78		90		99			

* For $p = 10$.

Table 3. Frequencies of the events $\{(\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{\alpha}_{33}) : |\hat{\alpha}_{11}| > 2N^{-1/2}\}$ *and* $\{(ri_1, ri_2, ri_3) : |ri_1| > 2N^{-1/2}\}$

		$N = 50$		$N = 150$		$N = 250$		Total	
Series I	IACF	20	20*	20	20*	20	20*	60	60*
	PACF	20		20		20		60	
Series II	IACF	0	1*	0	0*	1	0*	1	1*
	PACF	9		8		9		26	
Series III	IACF	20	20*	20	20*	20	20*	60	60*
	PACF	20		20		20		60	
Series IV	IACF	2	0*	0	0*	0	0*	2	0*
	PACF	3		2		1		6	
Series V	IACF	20	19*	20	20*	20	20*	60	59*
	PACF	20		20		20		60	
Series VI	IACF	0	0*	0	0*	0	0*	0	0*
	PACF	0		0		0		0	
Series VII	IACF	8	8*	20	20*	20	28*	48	56*
	PACF	12		20		20		52	
Series VIII	IACF	2	0*	0	0*	0	0*	2	0*
	PACF	3		0		0		3	
Total		72	68*	80	80*	81	88*		
		87		90		90			

* For $p = 10$.

model is close to the boundary of stationarity or has a characteristic equation with real roots and for $p = 10$.

Assessing the overall relative performance of the two functions in correctly determining the subset-order models, we find as evident

Table 4. Frequencies of the events $\{(\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{\alpha}_{33}): |\hat{\alpha}_{11}| \leq 2N^{-1/2}, |\hat{\alpha}_{22}| > 2N^{-1/2}, |\hat{\alpha}_{33}| \leq 2N^{-1/2}\}$ *and* $\{(ri_1, ri_2, ri_3): |ri_1| \leq 2N^{-1/2}, |ri_2| > 2N^{-1/2}, |ri_3| \leq 2N^{-1/2}\}$ for the subset-order models

		$N = 50$		$N = 150$		$N = 250$		Total	
Series II	IACF	18	18*	18	20*	15	19*	51	57*
	PACF	10		12		10		32	
Series IV	IACF	3	0*	4	1*	10	12*	17	13*
	PACF	3		3		11		17	
Series VI	IACF	20	20*	16	20*	19	19*	55	59*
	PACF	19		20		20		59	
Series VIII	IACF	0	0*	7	8*	7	5*	14	13*
	PACF	1		6		7		14	
Total		41	38*	45	49*	51	55*		
		33		41		48			

* For $p = 10$.

from Table 4 that the IACF is the better for all the models and for all the sample sizes.

Abraham and Ledolter (1984) base their use of $\pm 2N^{-1/2}$ for the assessment of both functions on the hypothesis of independence of observations. They also see the rationale for using $\pm 2N^{-1/2}\{1 + 2(ri_1^2 + \ldots + ri_{k-1}^2)\}^{1/2}$ for testing the significance of ri_k (on the basis of (3.3)), but rightly argue that its use would worsen its performance for determining the order of their models, all being full order. However, when the underlying subset model has some non-significant ri_j's,the IACF is likely to detect the significant non-zero lag coefficients more often than otherwise. In particular, for our subset-order models, the relative power of the IACF is higher with the non-null critical region than with $\pm 2N^{-1/2}$. The integral $\pm 2N^{-1/2}\{1+2(ri_1^2+\ldots+ri_{k-1}^2)\}^{1/2}$ is not adequately wider than the null critical region. An effect of this is that IACF is not much worse than the PACF in detecting the significance of the lag two coefficients. In addition, the IACF correctly suggests a zero value for the lag three coefficient more often.

It even supersedes the PACF in this sense, especially for $p = 10$ (see Table 5).

Table 5. Frequencies of the events $\{(\hat{\alpha}_{11}, \hat{\alpha}_{22}, \hat{\alpha}_{33}) : |\hat{\alpha}_{11}| \leq 2N^{-1/2}, |\hat{\alpha}_{22}| > 2N^{-1/2}, |\hat{\alpha}_{33}| \leq 2N^{-1/2}\}$ *and* $\{(ri_1, ri_2, ri_3) : |ri_1| \leq 2N^{-1/2}, |ri_2| > 2N^{-1/2}[1 + 2ri_1^2]^{1/2}, |ri_3| \leq 2N^{-1/2}[1 + 2(ri_1^2 + ri_2^2)]^{1/2}\}$ *for the following subset-order models*

		$N = 50$		$N = 150$		$N = 250$		Total	
Series II	IACF	19	20*	20*	20*	17	20*	56	60*
	PACF	10		12		10		32	
Series IV	IACF	2	0*	4	2*	11	12*	17	14*
	PACF	3		3		11		17	
Series VI	IACF	20	20*	20	20*	20	20*	60	60*
	PACF	19		20		20		59	
Series VIII	IACF	0	1*	7	8*	7	5*	14	14*
	PACF	1		6		7		14	
Total		41	41*	51	50*	55	57*		
		33		41		48			

* For $p = 10$.

6. Practical Examples

For the Yule-Walker (Y-W) approach, full order and subset order AR models were fitted using the above outlined order determination criteria. For Burg's maximum entropy (ME) approach we fitted only full-order models. For the criterion ϕ we used $c = 1.50$ and for FPEα $\alpha = 4$. Diagnostic checks were made by the use of Box-Jenkins (1976) portmanteau test statistic, which we denote by R.

6.1. Series A (Box and Jenkins (1976, pp. 525))

For the Y-W approach, BIC, ϕ, SIC, S and FPE4 recommend the full order AR(2)

$$X_t - 0.427X_{t-1} - 0.252X_{t-2} = \varepsilon_t, \qquad (6.1)$$

$$\hat{\sigma}^2 = 0.1002$$

with $R = 100.03$. The Box-Jenkins (1976) portmanteau test rejects the model. AIC, FPE, CAT_2, and CAT_3 choose the AR (7):

$$X_t - 0.373X_{t-1} - 0.197X_{t-2} - 0.020X_{t-3} - 0.014X_{t-4} + 0.015X_{t-5} - 0.062X_{t-6} - 0.156X_{t-7} = \varepsilon_t, \quad (6.2)$$

$$\hat{\sigma}^2 = 0.0950$$

which, with $R = 22.44$, is not discredited by the portmanteau test.

For the ME approach, BIC, ϕ, SIC, and FPE4 choose an AR(2) very close to (6.1) and for which R is also significant. FPE, AIC, S, CAT_2, and CAT_3, however, choose an AR(7) very close to (6.2) and which is also recommended by the R-test.

Applying the subset AR modelling algorithm of Haggan and Oyetunji (1984), for a maximum lag of 15 gave the following subset models: BIC, S, SIC, ϕ, and FPE4 recommend the model:

$$X_t - 0.381X_{t-1} - 0.216X_{t-2} - 0.188X_{t-7} = \varepsilon_t, \quad (6.3)$$

$$\hat{\sigma}^2 = 0.0955,$$

$$R = 24.31.$$

CAT_2 and CAT_3 pick (6.1). AIC chooses:

$$X_t - 0.388X_{t-1} - 0.220X_{t-2} - 0.174X_{t-7} - 0.126X_{t-14} + 0.122X_{t-15} = \varepsilon_t, \hat{\sigma}^2 = 0.0934 \quad (6.4)$$

$$R = 20.15.$$

FPE selects the full order AR(15). Both (6.3) and (6.4) were found adequate by the R-test. To confirm the result of the R-test, Etuk (1987) has shown that the models (6.2), (6.3), and (6.4) have spectra which closely agree with an estimate of the raw spectrum; (6.1) does not. The model (6.3) is the most adequate used here since it is the most parsimonious amongst the adequate models (6.2), (6.3), and (6.4). Cleveland (1972) has also suggested the model

$$X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_7 X_{t-7} = \varepsilon_t.$$

Table 6. A comparison of the PACF and IACF for series A

Lag	Y-W PACF	Y-W AR(10) IACF	Burg's AR(10) IACF	Burg's PACF
1	0.57	−0.31	−0.30	0.57
2	0.25	−0.16	−0.15	0.25
3	0.07	−0.02	−0.02	0.08
4	0.07	−0.01	−0.02	0.09
5	0.07	0.01	0.02	0.07
6	0.12	−0.06	−0.07	0.14
7	0.15	−0.14	−0.16	0.19
8	−0.03	0.02	0.03	−0.04
9	0.01	−0.02	−0.02	0.01
10	−0.02	0.01	0.01	−0.01

With $2/\sqrt{N} = 0.14$, we see that the IACF not only correctly recommends an order of 7, but also better suggests zero parameter values for lags 3, 4, 5, and 6, especially for lag 6.

6.2. *Canadian lynx numbers (1821 – 1934) (Campbell and Walker (1977, pp. 430))*

The well-analyzed logarithm transformation has been used. For the Y-W method for full-order modelling, BIC, ϕ, and SIC pick the order 2 which though recommended by the *R*-test has a spectrum which does not tally with the estimated raw spectrum. However, the AR(11) selected by FPE, AIC, S, CAT_2, CAT_3, and FPE4 is found adequate by both diagnostic checking criteria. For the ME technique, BIC and ϕ pick an order 2 also found to be inadequate by the spectrum test, although its *R* value is nonsignificant. SIC, S, CAT_2, CAT_3, and FPE4 pick the AR(11) found to be adequate by both tests. FPE and AIC pick the likely overparametrized AR(12). For the full-order modelling order 11 has been found to be the best (Etuk (1987)).

For subset-order modelling, BIC, ϕ, SIC, and FPE4 recommend an order 11 with significant lags 1, 2, 4, 10, and 11. CAT_2 and CAT_3 select the model with significant lags 1, 2, and 4. FPE chooses the full order AR(15). Etuk (1987) has shown that the subset model with lags 1, 2, 4, 10, and 11 is the best model.

Table 7 gives the PACF and IACF. With $N=114$, $2/\sqrt{N} = 0.19$. Therefore, the PACF suggests an order of 4, 7, or 11. The IACF, however, suggests an order of 1. The PACF is significant at lags 1, 2, 4, 7, 10, and 11 which tallies better with the identified model. Evidence here is therefore in favour of the PACF.

Table 7. A comparison of the PACF with the IACF for the log transform of the lynx data

Lag	Y–W PACF	Y–W AR(15) IACF	Burg's AR(15) IACF	Burg's PACF
1	0.79	−0.40	−0.39	0.79
2	−0.72	0.17	0.17	−0.74
3	−0.14	−0.07	−0.09	−0.12
4	−0.21	0.10	0.12	−0.21
5	0.12	−0.05	−0.06	0.14
6	0.08	0.05	0.06	0.07
7	0.21	−0.03	−0.03	0.23
8	0.12	0.03	0.03	0.13
9	0.10	−0.06	−0.06	0.12
10	−0.19	−0.04	−0.05	−0.22
11	−0.31	0.06	0.07	−0.35
12	−0.10	0.08	0.06	−0.13
13	0.10	−0.04	−0.00	0.05
14	−0.04	−0.00	−0.03	−0.01
15	−0.02	0.01	0.01	−0.04

6.3. *Wolfer's sunspot series (1700 – 1955)*

Data on sunspots are available from 1700 onwards (see Waldmeier (1961)). We used 256 values from 1700 to 1955.

For the Y-W approach for full-order model selection FPE, AIC, and CAT_3 recommend an AR order of 9; FPE4, ϕ, and S recommend 8; SIC recommends 3, CAT_2 zero, and BIC 2. For the ME technique, FPE and AIC pick an AR(18); BIC, SIC, and ϕ pick AR(8); CAT_3 and FPE4 choose the AR(9); CAT_2 selects the AR(0) and S the AR(10).

For subset modelling, BIC, SIC, ϕ, and FPE4 recommend a model with lags 1, 2, and 9. AIC pick the lags 1, 2, 3, 4, 5, and 9. CAT_3 choose the lags 1, 2, and 3; S choose the lags 1, 2, 3, and 9. The BIC subset model has been shown to be the most adequate model (Etuk (1987)).

The critical value for both functions is 0.125. With PACF significant lags are 1, 2, 3, 6, 7, 8, and 18, so that the suggested orders are 3, 8, and 18. The IACF indicates an order of 2 (see Table 8).

We therefore observe an agreement of the PACF with BIC, SIC, and ϕ in choosing an order of 8, with AIC in choosing an order 18. The IACF agrees with BIC in the choice of order 2. The two diagnostic checking methods used here do not discredit an order of 8. The order of 18 is therefore an overestimation. The orders 2 and 3 are too low for the model.

Thus we find that PACF is the better model identifier.

Table 8. Comparison of IACF and PACF for sunspot series

Lag	Y-W PACF	Y-W AR(15) IACF	Burg's AR(20) IACF	Burg's PACF
1	0.81	0.41	0.41	0.82
2	−0.66	0.15	0.14	−0.67
3	−0.15	0.07	0.07	−0.14
4	0.04	−0.08	−0.09	0.05
5	−0.07	0.08	0.07	−0.07
6	0.17	−0.05	−0.03	0.18
7	0.14	0.03	0.01	0.19
8	0.22	−0.04	−0.04	0.23
9	0.10	−0.05	−0.05	0.12
10	0.01	0.04	0.03	0.03
11	0.07	−0.06	−0.06	0.07
12	−0.07	0.02	0.02	−0.07
13	0.00	0.04	0.05	0.00
14	0.06	−0.05	−0.06	0.07
15	−0.09	0.03	0.04	−0.10
16	−0.06	−0.01	−0.00	−0.07
17	−0.10	−0.02	−0.02	−0.11
18	−0.13	0.05	0.06	−0.16
19	−0.01	−0.02	−0.02	−0.00
20	0.03	0.01	0.01	−0.02

7. Conclusion

Evidence here is not conclusive as to the better criterion. However, we have seen that the PACF is definitely the better in identifying non-zero lag coefficients and determining order. Invariably, the PACF outperforms the IACF for full-order AR models. The reason for this is that though, as observed by Abraham and Ledolter (1984), chances are that $|\hat{\alpha}_{jj}| > |\hat{\gamma}_{jj}| \; \forall j = 1, 2, \ldots$, the variance of the PACF is the smaller, for any given lag. For the same reason, the IACF more often correctly detects zero coefficients, especially where they are intervening.

We have observed that the PACF outperforms IACF for two of the three real series we used. Even for the lynx data for which a subset AR model is suggestive, PACF fares the better. However, our Monte Carlo study shows that the IACF is generally the better for subset-order modelling.

We have also noticed a tendency of overestimation with the PACF and of underestimation with the IACF. Exclusive preference of one to the other may therefore not be advisable.

The IACF and PACF of an AR model is the autocorrelation function (ACF) and inverse partial autocorrelation function (IPACF) respectively of the inverse moving average (MA) model. Comparision of the IACF and PACF in AR model identification is tantamount to that of the ACF and IPACF in MA modelling. Therefore, we suggest the application of all four functions: ACF, IACF, PACF, and IPACF in autoregressive moving average (ARMA) modelling.

8. References

Abraham, B. and Ledolter, J. (1984): A Note on Inverse Autocorrelations. Biometrika, 71, pp. 609–614.

Akaike, H. (1969): Fitting Autoregressive Models for Prediction. Annals of the Institute of Statistical Mathematics, 21, pp. 243–247.

Akaike, H. (1977). On Entropy Maximization Principle. Proceedings of the Symposium on Applied Statistics. (P.R. Krishnaiah, ed.) Amsterdam, North – Holland.

Andersen, N. (1974): On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis. Geophysics, 39, (1), pp. 69–72.

Bhansali, R.J. (1975): Fitting Time Series Models in the Frequency Domain. Bulletin of the International Statistical Institute, 46, Book 3, pp. 98–104.

Bhansali, R.J. (1983): A Simulation Study of Autoregressive and Window Estimators of the Inverse Correlation Function. Applied Statistics, 32, pp. 141–149.

Bhansali, R.J. and Downham, D.Y. (1977): Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's FPE criterion. Biometrika, 64, pp. 547–551.

Box, G.E.P. and Jenkins, G.M. (1976): Time Series Analysis Forecasting and Control. Holden-Day, San Francisco.

Campbell, M.J. and Walker, A.M. (1977). A Survey of Statistical Work on the Mackenzie River Series of Annual Canadian Lynx Trappings for the Years 1821–1934 and a New Analysis. Journal of the Royal Statistical Society, Series A, 140, pp. 411–431.

Chatfield, C. (1979): Inverse Autocorrelations. Journal of the Royal Statistical Society, Series A, 142, pp. 363–377.

Cleveland, W.S. (1972): The Inverse Autocorrelations of a Time Series and Their Applications. Technometrics, 14, pp. 277–297.

Etuk, E.H. (1987): On the Selection of Autoregressive Moving Average Models. Ph.D. Thesis. Department of Statistics, University of Ibadan, Nigeria.

Haggan V. and Oyetunji, O.B. (1984): On the Selection of Subset Autoregressive Time Series Models. Journal of Time Series Analysis, (2), pp. 103–113.

Hannan, E.J. and Quinn, B.G. (1979): The Determination of the Order of an Autoregression. Journal of the Royal Statistical Society, Series B, 41, pp. 190–195.

Hipel, K.W., McLeod, A.I., and Lennox, W.C. (1977): Advances in Box-Jenkins Modelling 1. Water Resources Research, 13, pp. 567–575.

Hosking, J.R.M. (1980): The Asymptotic Distribution of the Sample Inverse Autocorrelation of an Autoregressive- moving

Average Process. Biometrika, 67, pp. 223–226.

McClave, J. (1978): Estimating the Order of Moving Average Models: The Max χ^2 Method. Communications in Statistics, A (73), pp. 259–276.

McLeod, A.I., Hipel, K.W., and Lennox, W.C. (1977): Advances in Box-Jenkins Modelling 2. Water Resources Research, 13, pp. 577–586.

Oyetunji, O.B. (1985): Inverse Autocorrelations and Moving Average Time Series Modelling. Journal of Official Statistics, 1, pp. 315–322.

Parzen, E. (1977): Non-parametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for "White Noise". Statistical Science Division Report No. 47. State University of New York at Buffalo.

Priestley, M.B. (1981): Spectral Analysis and Time Series. Academic Press, London.

Schwarz, G. (1978): Estimating the Dimension of a Model. Annals of Statistics, 6, pp. 461–464.

Shibata, R. (1980): Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear process. Annals of Statistics, 8, pp. 147–164.

Tong, H. (1977): On Some Alternative Approaches to Autoregressive Order Determination. Technical Report 80, Department of Mathematics, University of Manchester, Institute of Science and Technique, U.K.

Waldmeier, M. (1961): The Sun-Spot Activity in the Years 1610–1960. Schulthess, Zurich.

Yule, G.U. (1927): On a Method of Investigating the Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. Philosophical Transactions of the Royal Society London (A), pp. 226–267.

Received September 1987
Revised April 1988

Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 125–140

Classifying and Comparing Spatial Relations of Computerized Maps for Feature Matching Applications

Alan Saalfeld[1]

Abstract: Modern computerized maps either contain digital information on spatial relations, such as adjacency relations, shape, network patterns, and measures of position and distance of features, or they permit derivation of that information from the feature data that they do contain. Such spatial attributes lend themselves to computerized statistical analysis much like any other data. Comparative data analysis of spatial relations is possible when two map files are known to cover the same area. In this case, spatial characteristics alone may be used to establish linkages between many of the feature records of the two files. This paper presents examples of some spatial measures of distance and local configuration that were used to develop an automated feature matching system at the U.S. Bureau of the Census. For a particular sample pair of maps, global summaries and spatial depictions of distance and configuration measures are presented; and some additional uses for the measures are suggested.

Key words: Computerized maps; map distortion; automated cartography; feature matching; record linkage; configuration; conflation; spider function.

1. Introduction

1.1. Background

Conflation is the consolidation or merging of two map representations of the same region into a third composite conflated map. Recently the U.S. Bureau of the Census has begun consolidating or conflating pairs of digital (computerized) map files of the same region to measure

[1] Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, U.S.A.

Acknowledgments: The author wishes to cite the outstanding computer graphics support provided throughout this research work by Maureen Lynch of the Statistical Research Division of the U.S. Bureau of the Census.

and improve the quality of the bureau's digital maps. A second set of digital maps for the entire country is being provided by the United States Geological Survey (USGS) for the U.S. Bureau of the Census to use with its own metropolitan map files for comparative updating of both sets of maps. The second set of USGS digital maps was created by mechanically scanning line drawings of road and water networks and thus contains only spatial information about line segments and their intersections. It does not contain any name or attribute information. Thus, only comparisons involving line segments, their locations and locations of their intersections, and derived spatial measures are possible. All of the work in this paper, therefore, treats a map as nothing more than a plane line graph or network.

In the past, measures of similarity and differences of linear features of maps, primarily of paper maps, were not quantitative or even fully quantifiable; and this limitation made the comparative analysis of maps quite subjective and nonnumerical. Often differences and discrepancies were merely noted or listed; and there was no readily understood measure of map similarity. The digital map file, on the other hand, is by its very nature considerably more amenable to numerical analysis, and its format invites computer analysis.

1.2. Scope of this paper

Now it is not only feasible and informative to quantify and analyze individual linear feature similarities and differences; it is also useful to develop concrete numerical measures and graphic displays of regional and global similarity to establish statistically that two maps, or specific significant regions within the two maps, are piece by piece the same. A challenging problem is to find a local or regional numerical signature that can be used to block or group together feature records to limit or localize a search for matches. Finding such a blocking algorithm, typically a critical component to any record linkage system, is currently the principal obstacle to fully automating the feature record matching subsystem of the conflation system.

The aim of this paper is to present some initial attempts at quantifying map similarities and differences when the maps consist exclusively of spatial information. The paper outlines approaches to analysis of those differences and similarities; it does not contain extensive empirical justification for those approaches. This last constraint is due in part to the limited available data. Although the Bureau of the Census will eventually have to conflate over 5 500 map pairs (each map covering approximately 50 square miles), only three such map pairs were made available for this research. While the results of the initial quantification measures are encouraging in the few examples to which they have been applied, it is necessary to note that the measures themselves are only a few of many possible measures; and the observations based on three map sets illustrate the potential for, rather than prove, the measures' effectiveness.

2. Conflation and Automated Feature Matching

A cartographer, in order to compile two maps of the same region and produce a third new map, uses numerous visual clues and cues to match features of one map to features of the other; and, when he/she is convinced of a match, he/she extracts a single common feature from the two maps. After a cartographer has matched features on the two maps, a statistical analysis of the numerical properties of the matched and unmatched features may be performed. The resulting analysis yields information on the numerical characteristics of the cartographer's matching operation or matching algorithm. The resulting analysis, in turn, may be used to develop a rule-based system and to drive an automatic statistical matching procedure, which can then replicate the cartographer's results and, thus, auto-

mate the map conflation process. Due to the need for uniform processing and the large number of map files to be processed, the final production system for computerized matching and merging of two map files should be as fully automated as possible.

At the U.S. Bureau of the Census, to assess various rules for matching, a semi-automatic interactive color graphics prototype conflation system has been implemented on a Tektronix 4125B Workstation (Lynch and Saalfeld (1985)). A computer operator uses the system to manipulate map images and to classify street intersections and street segments as matches or nonmatches.

The system is semi-automatic in that it has been programmed to initiate the feature-matching detection of a cartographer by applying various matching criteria and then prompting the operator with its findings. The position and configuration characteristics of the map features being compared serve as criteria. After the computer locates a likely match based on the matching criteria, the operator needs only to verify or reject the proposed match. The use of color to distinguish between the two maps and to distinguish features that have already been classified as matches or nonmatches has also facilitated operator decision-making procedures. After matches have been confirmed, fast rubber-sheeting[2] algorithms are used to align the maps, thereby permitting effective immediate visual verification of matching decisions. The most valuable element of the color graphics and image alignment approach has been the ease and accuracy of assessing whether or not a match was made correctly. The currently used matching and alignment procedure is iterative; with each iteration, it brings more and more matched feature pairs into exact alignment, moves matchable pairs closer and closer together, and moves pairs which do not match farther and farther apart (Saalfeld (1985)).

[2] Rubber-sheeting refers to transformations of the plane or rectangular subregion that preserve topological invariants. Piecewise linear homeomorphisms are elementary instances of topology-preserving or rubber-sheeting transformations.

3. Differences Within and Between Maps

3.1. Measures of feature position or location

This study of map similarities and differences focuses on street intersections and their configuration and location. Intersection locations are stored by their coordinates; and as one would expect, the intersections are not clustered in space, but are fairly evenly distributed in the plane, as shown in Fig. 1B.

The average Euclidean distance from any intersection to its nearest neighbor intersection on the same map is large compared to the average amount of local distortion on different maps; and this fact makes an image alignment approach effective.

Distortion is most easily analyzed through overlay techniques. Alignment may be achieved through elementary transformations called rubber-sheeting functions that relocate key points of one or both maps on top of corresponding points and move other points of the maps proportionately. The tranformations used in the Census Bureau system are piecewise linear homeomorphic (PLH) functions defined on a triangulation of the map space or spaces (Griffin and White (1985)).

Others have used smooth functions such as bivariate quintics, again defined on triangulations, (Lupien and Moreland (1987)) for their rubber-sheeting alignment.

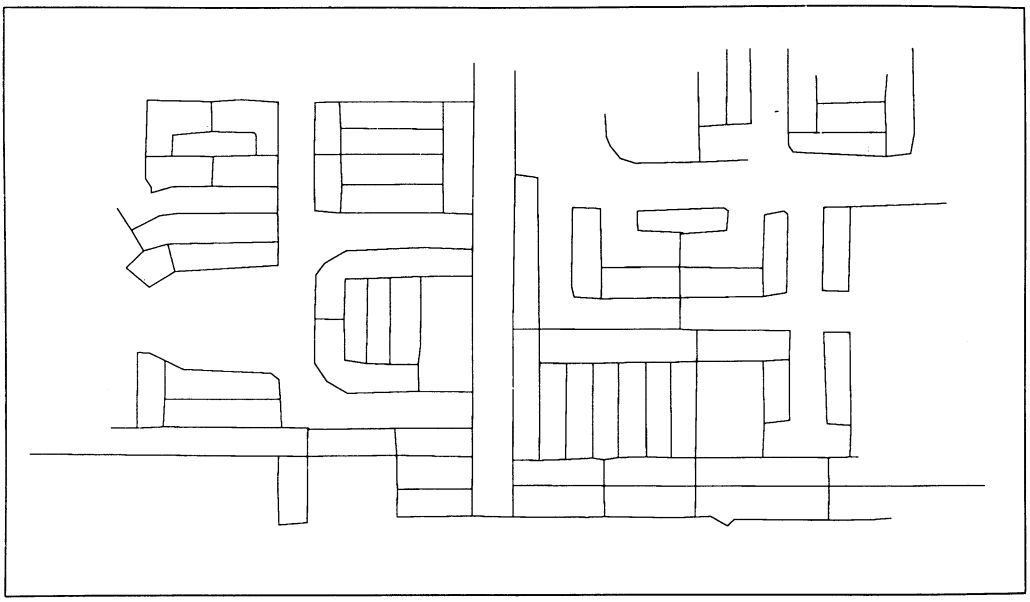

Fig. 1A. USGS map of part of Fort Myers, FL

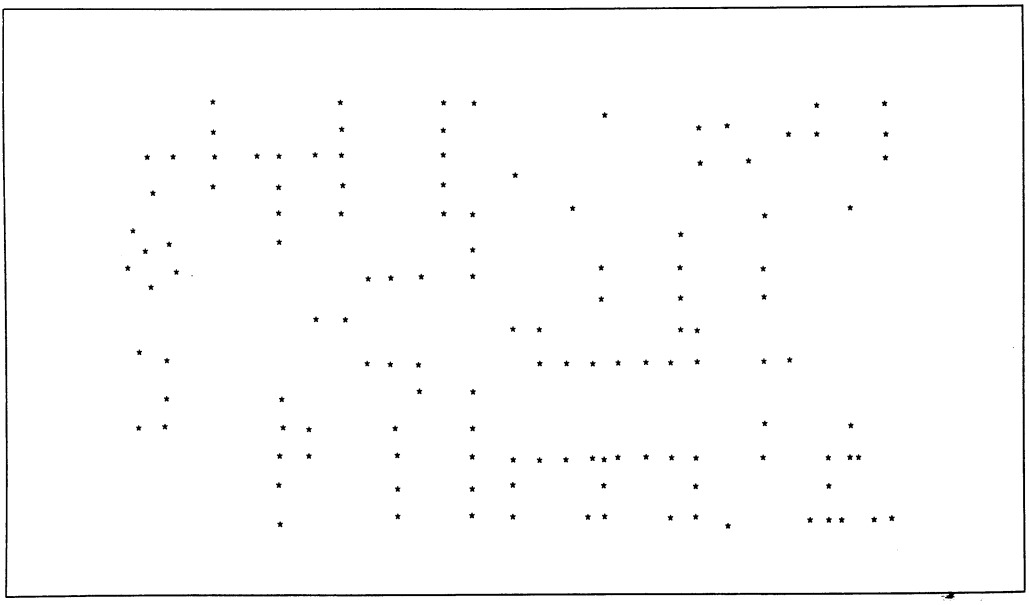

Fig. 1B. Street intersection point distribution for the same map as in Fig. 1A

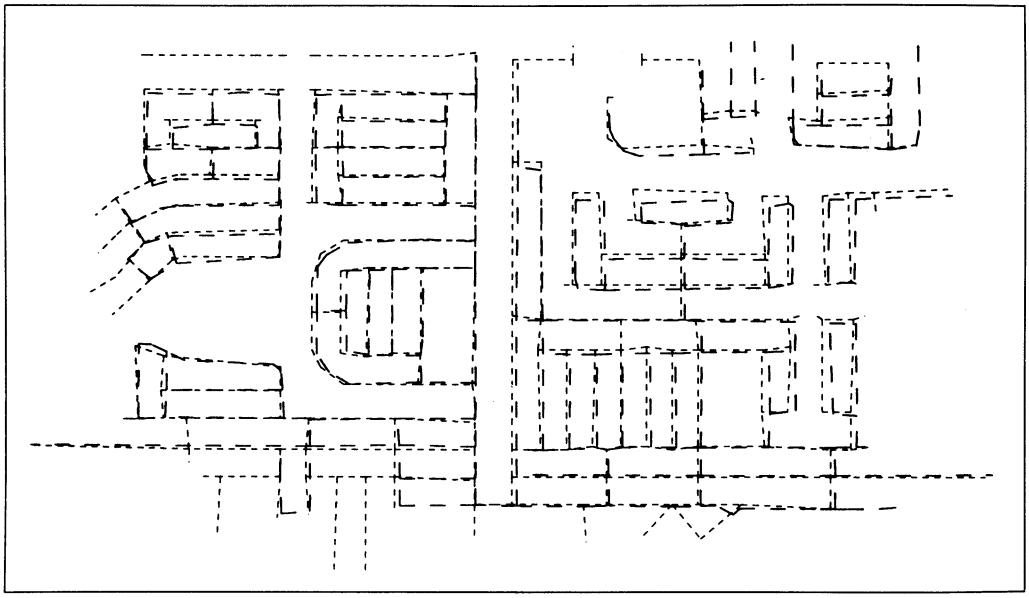

Fig. 2A. Overlay of two map sections of Fort Myers, FL

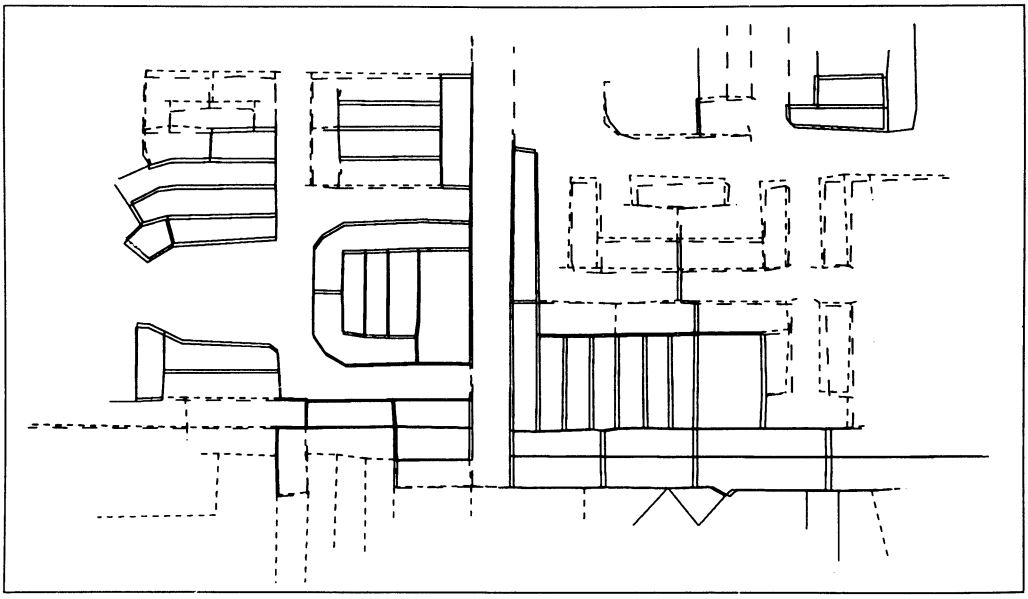

Fig. 2B. Matched and aligned sections of the same area

Figures 2A and 2B suggest that a good initial alignment achieved with PLH transformations can bring nearly all matchable pairs into proximity. The proximity condition is so strong that being a nearest street intersection on the other map almost becomes a necessary (but not sufficient) condition for intersection matchability.

Exploratory studies of distortion (Lupien and Moreland (1987)) have displayed as elevation the displacement in each coordinate direction between maps to produce distortion surfaces such as the following:

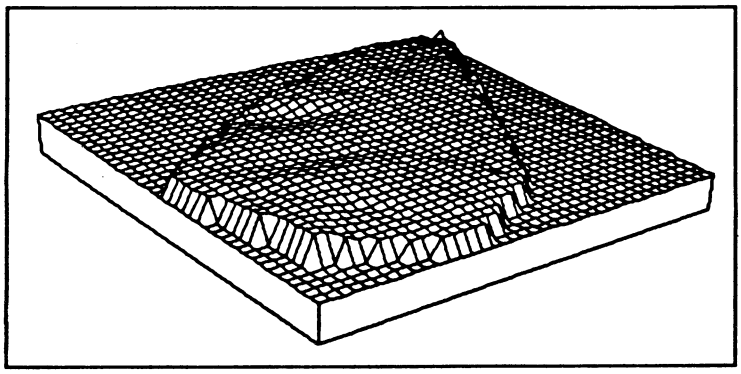

*Fig. 3A. 50 link distortion surface for X coordinate**

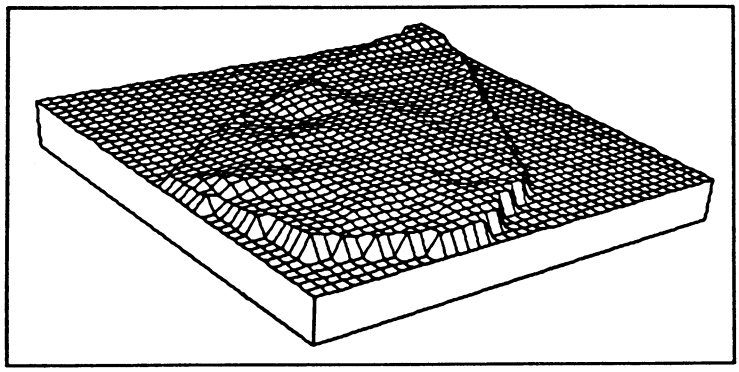

*Fig. 3B. 50 link distortion surface for Y coordinate**

*Figures 3A and 3B reproduced with authors' permission.

Available rubber-sheeting techniques have no difficulty aligning maps of different scales and orientations. Distortion surfaces measure the amount of movement required for that alignment. The mean slope of each distortion surface, for example, would reflect the overall scale difference in each of the respective coordinates. Orientation change has a similarly predictable and detectable effect on the distortion surfaces.

In general error theory, two types of false classifications may occur. A map feature may be labelled incorrectly as having a match when indeed it does not (false positive); or a feature may be judged incorrectly not to match any feature when it has a true pairing (false negative). In matching theory a third type of error occurs when a feature is judged correctly to have a match, but the wrong correspondent is judged to be the matching element. This type of error is called mismatch. The iterative matching procedure used with the conflation system identifies new matches at each stage and does not label nonmatches as such until the final stage. False negatives are a residual and do not present a problem at an intermediate iteration. False positive errors and mismatches are less desirable and less managable than false negatives because they may precipitate additional errors at subsequent iterations, and at no point in the iteration procedure is there an unmatching capability for correcting false positives and mismatches.

The Euclidean distance between potential matches after initial alignment is an excellent measure for controlling both mismatch and false negative errors. For one particular test map of Fort Myers, FL, Table 1 shows the distribution of instances of distance ranges from matchable points (points for which a match was found and visually verified) on the Census map to their matched or paired points on the USGS map (column 2), and from the same matchable points on the Census map to their nearest nonmatching neighbors on the USGS map (column 3).

Table 1. Distribution of distances from matchable points to their matches and nearest nonmatches (after initial PLH alignment)*

Distance range (meters)	Number of matchable points whose matching pair is within range	Number of matchable points whose nearest nonmatch is within range
0– 5	162	–
5– 10	359	–
10– 15	272	4
15– 20	132	8
20– 25	70	14
25– 30	19	25
30– 40	13	54
40– 50	3	90
50– 60	2	227
60– 70	–	302
70– 80	1	134
80–100	–	86
100–200	1	82
200–400	–	8
400 and above	–	–
All distances	1 034	1 034

	Mean distance	Range	Standard deviation
To matching point	11.45	112.25	7.75
To nearest nonmatch	66.68	278.89	28.55

*PLH alignment uses 32 local alignments and 66 triangles.

The initial alignment used to produce Table 1 was accomplished through hardware and software image manipulation of Census and USGS maps. First the Census map was subdivided into 32 equal-sized rectangular pieces. Each rectangular piece could be moved anywhere on the screen by the operator. Using the entire USGS map as background, the operator positioned each small census rectangle to produce the best possible visual alignment near each small rectangle's centroid. The movement that had been required to position the 32 centroids was recorded and averaged locally (using PLH functions on a triangulation of the Census map) to rubber-sheet the Census map and recompute all of its coordinates (Saalfeld (1985)).

The cumulative relative frequencies shown in Figures 4A and 4B, which summarize Table 1, support the idea that, after initial map alignment, nearest neighbor pairs are excellent candidates for matching.

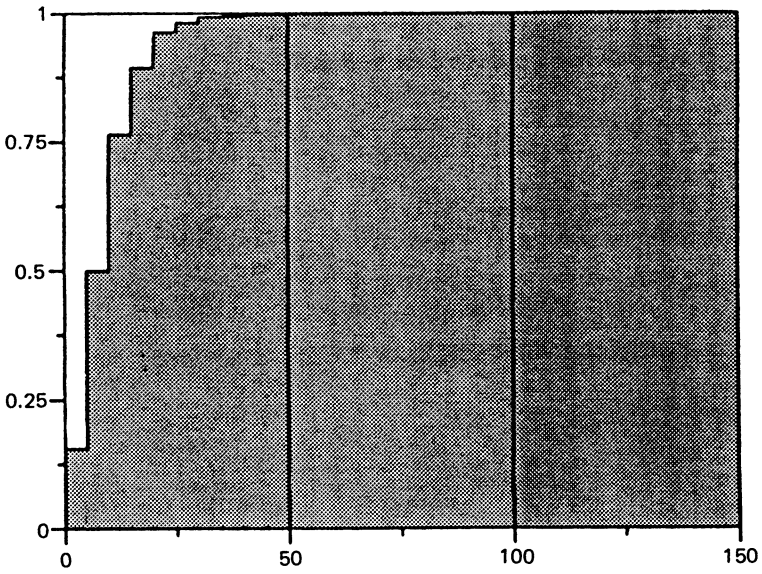

Fig. 4A. Fraction of matchable points whose matching point is within the indicated distance of the point

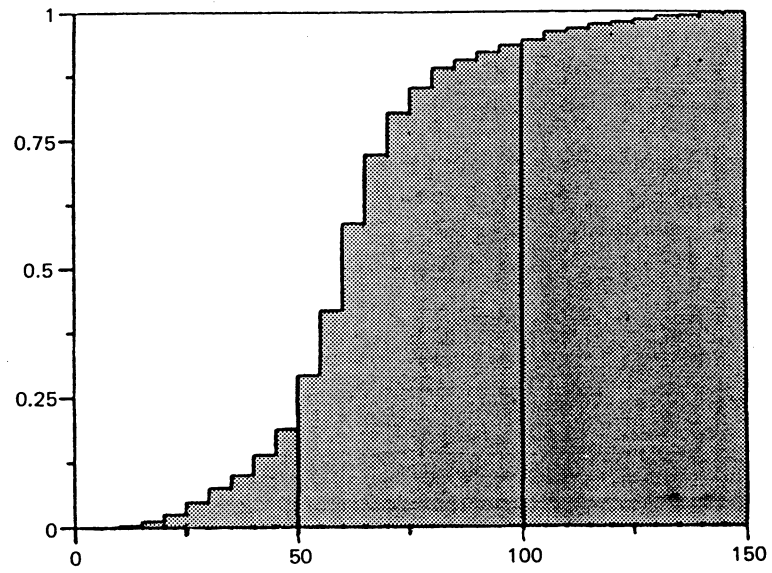

Fig. 4B. Fraction of matchable points whose nearest non-matching point is within the indicated distance of the point

Nearness alone will not suffice for matching. Nonetheless, distance tolerances may be used for estimating both mismatch and false negative error types and reducing the one type or the other. In the Fort Myers map, for example, if the threshold for matching is set at 20 meters, (that is, no matches are accepted unless the candidate pairs are within 20 meters of each other), then the measured probability of omitting a match (false negative) is 11%, and the probability of mismatching a matchable point is 1%. By decreasing the threshold, mismatches may be reduced further. However, the increase in false negatives will require additional iterations of the file processing; and the threshold may even need to be relaxed in the final iterations to detect all matches.

3.2. Measures of configuration

The remainder of this paper focuses on other match criteria tests to supplement nearest neighbor tests. To facilitate quantitative comparisons of intersection patterns, the configurations are assigned numerical summary values and are grouped according to those values. The coding scheme reflects similarities of patterns through the assignment of nearly equal summary values when the intersection configurations themselves are nearly identical (Rosen and Saalfeld (1985)). These additional criteria utilize the following numerical measures of local configuration.

3.2.1. The degree of an intersection

The number of streets emanating from an intersection is called the *degree* of the intersection. The degree provides a good measure on which to match intersections if it is unique or locally unique (e.g., the only intersection in the neighborhood with seven streets coming into it.)

3.2.2. The spider function of an intersection

The street pattern at an intersection (that is, the emanating rays) has infinitely many possibilities for street directions. To simplify the possibilities, the number of directions was reduced to eight sectors. The eight sectors correspond to 45° pie slices centered upon the principal directions of north, northeast, east, southeast, south, southwest, west, and northwest. The eight sectors in counter-clockwise order are assigned consecutive bit positions (from right to left) in an eight-bit binary number, and the bit for a given sector is changed from "0" to "1" if and only if there is a street in that sector. The resulting number has been named descriptively the spider function of the intersection. With this function, an integer between 1 and 2^8–1 describes the street pattern of the intersection. The binary number 01010101 (which is the decimal 85 and hexadecimal 55) represents the typical four-street north–south–east–west intersection, for example. The street pattern is assumed to have at most one street in each of the eight sectors. (If more than one street occurs in any sector, the spider function may be given a special value or it may simply ignore the extra street. Limited experience suggests that ignoring the extra street will not adversely affect our matching procedure since (1) two streets in the same sector are very rare, and (2) matching is allowed if street configurations are only similar – e.g., "off by one" – and not identical.) Intersection patterns whose difference is a power of two are usually "close" in one of two geometric senses: either one pattern is missing a single street, but agrees everywhere else; or else one street is shifted, off by a single sector. By comparing the *degree* of an intersection as well as the spider function, the U.S. Bureau of the Census has developed several simple measures of nearness of configuration.

The representation of the spider function value as a hexadecimal (base 16) integer has additional nice properties.

1. The spider function value is always a two-digit number.
2. Each digit describes the street directional behavior in a four-sector band constituting a semicircular region.

3. A digit K in the second (units) position describes the same configuration as the same digit K would describe in the first (sixteens) position except for a rotation of 180° (see Fig. 5).
4. The configuration with hexadecimal digits NM is the 180° rotation of the configuration with hexadecimal representation MN, the number with digits M and N transposed.
5. Numbers with repeated digits KK and only those numbers have all streets continuing straight through the intersection.

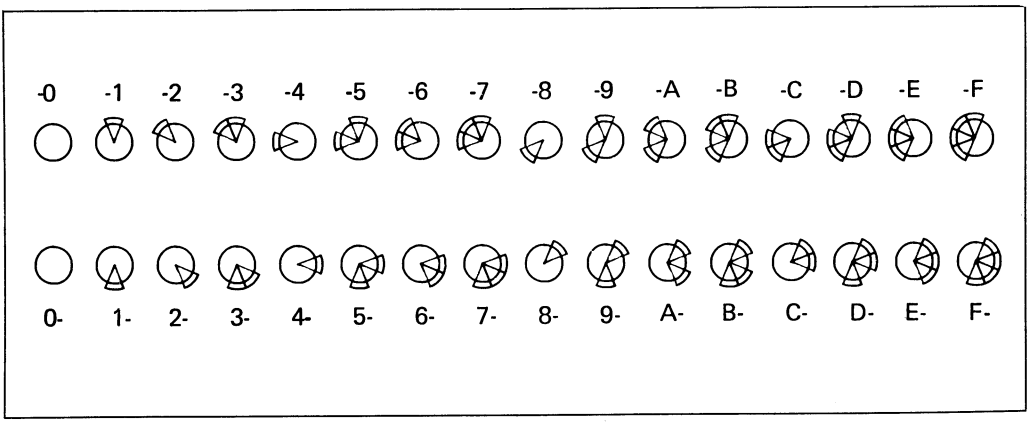

Fig. 5. Hexadecimal and sector patterns for spider function

3.3. Summary statistics on global configuration

3.3.1. Spider function tables

A frequency distribution of spider function values for a map may be organized in a sixteen-by-sixteen table whose columns correspond to second (units) digit values and whose rows correspond to first (or sixteens) digit possibilities in the hexadecimal representation. In a highly urbanized area, for example, the frequency of the hexadecimal number 55, representing the north–east–south–west intersections, would be very large, and could help distinguish between urban and other areas. More generally, the frequency table establishes a kind of signature for the street network; and parts of the table, such as the diagonal, have special meaning. (The principal diagonal of the table is comprised precisely of those intersections all of whose streets continue straight through the intersection.)

Two tables (one for the USGS map and one for the Census map) showing the distribution of spider function values for all map intersections for the 25 square mile Fort Myers area are given below. Such tables can orient an initial exploratory data analysis of intersection patterns of the area. After viewing the tables, one may display, in the plane, all of those intersection points having a particular spider value (or a range of related spider values) and then proceed to apply pattern recognition techniques to the pattern, as is illustrated below.

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	0	36	7	2	38	29	2	-	5	2	1	1	5	6	1	-
1-	35	23	1	19	28	280	11	-	3	16	2	-	2	-	-	-
2-	9	3	4	6	1	9	10	-	3	4	9	1	1	-	1	-
3-	1	13	4	-	10	2	-	-	1	-	-	-	-	-	-	-
4-	28	22	4	14	22	315	12	-	3	18	7	-	6	4	-	-
5-	40	273	10	-	304	225	3	-	10	3	-	-	-	-	-	-
6-	6	10	4	1	4	2	4	-	-	-	5	-	1	-	-	-
7-	1	-	-	1	-	-	2	-	-	-	-	-	-	-	-	-
8-	8	-	1	2	3	13	2	-	3	10	21	-	3	-	-	-
9-	-	19	2	-	17	-	-	-	21	2	2	-	-	-	-	-
A-	2	2	10	-	10	-	2	-	29	3	9	-	-	-	-	-
B-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	4	1	2	-	11	-	-	-	4	-	2	-	-	-	-	-
D-	5	1	1	-	2	-	-	-	1	-	1	-	-	-	-	-
E-	1	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2A. Spider function distribution for USGS intersections

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	-	93	22	-	80	16	-	-	21	1	1	-	-	6	-	-
1-	78	23	4	10	8	188	4	-	-	11	1	-	-	2	-	-
2-	14	2	5	4	4	12	7	-	1	1	6	2	3	-	-	1
3-	-	8	3	1	6	2	-	-	-	-	-	-	-	-	-	-
4-	77	10	5	9	31	204	13	1	5	21	5	1	14	1	-	-
5-	10	179	10	3	204	154	4	-	13	5	1	-	-	-	-	-
6-	3	6	3	-	10	7	3	-	2	-	5	-	1	-	-	-
7-	-	2	2	1	-	-	1	-	-	-	-	-	-	-	-	-
8-	12	1	2	1	3	8	2	-	2	1	17	-	9	-	-	-
9-	2	13	10	-	14	2	-	-	28	1	2	-	-	-	-	-
A-	2	5	8	-	8	1	-	-	25	3	12	-	4	-	-	-
B-	4	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
C-	2	-	5	-	14	-	-	-	5	-	-	-	1	-	-	-
D-	7	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-
E-	1	-	-	-	1	-	-	-	-	-	1	-	-	-	-	-
F-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2B. Spider function distribution for Census intersections

As an illustration of exploratory analysis that can be applied to the above tables, notice that the total number of intersections for the USGS map is far greater than the total for the Census map. This difference is due to the greater extent or coverage of the USGS map. The Census map merely covers a subregion of the USGS map. Nevertheless, cell percentages are very similar, indicating that the distribution of intersections by configuration types is the same. Moreover, the anomaly of having fewer "55" or north–south–east–west intersections than any type of "T" intersection: 15, 51, 45, and 54, is apparent in both tables. The prevalence of "T" intersections in the Fort Myers area is due to frequent water inlets that result in numerous natural road barriers. It is indeed a signature or identifying characteristic for the area.

Since the occurrences are linked to spatial position, the tables shown above could further be decomposed according to subareas or subregions of the map. Although the total number of entries would decrease, the entries present would then reflect more accurately local characteristics of the chosen subarea of the street network.

3.3.2. Spider displays as point patterns

After the spider function tables are compiled, one may choose to display as point patterns only those intersections whose occurrences in the spider function tables are judged extraordinary. One may look at rare occurrences such as the unique "6C" intersection appearing on both maps; or one may draw all "15 T" intersections to try to determine why they are so frequent. The second option is illustrated in the figures below as a filtering operation. In the first set of figures the entire range of spider function values in a subregion are plotted in their intersection locations. In the other sets only those intersections with particular spider function values are plotted.

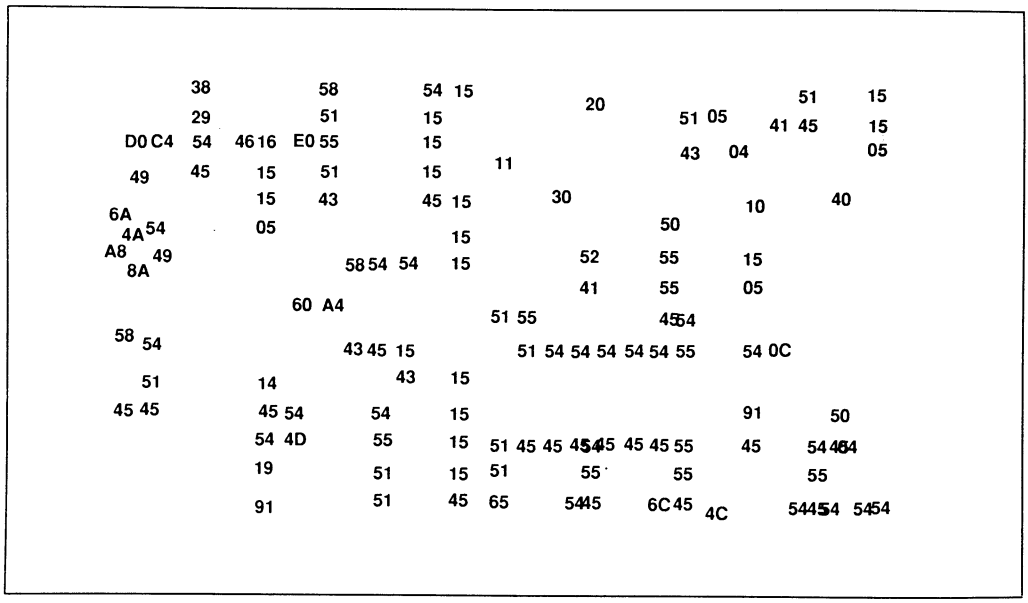

Fig. 6A. All spider function values in a subarea of USGS map

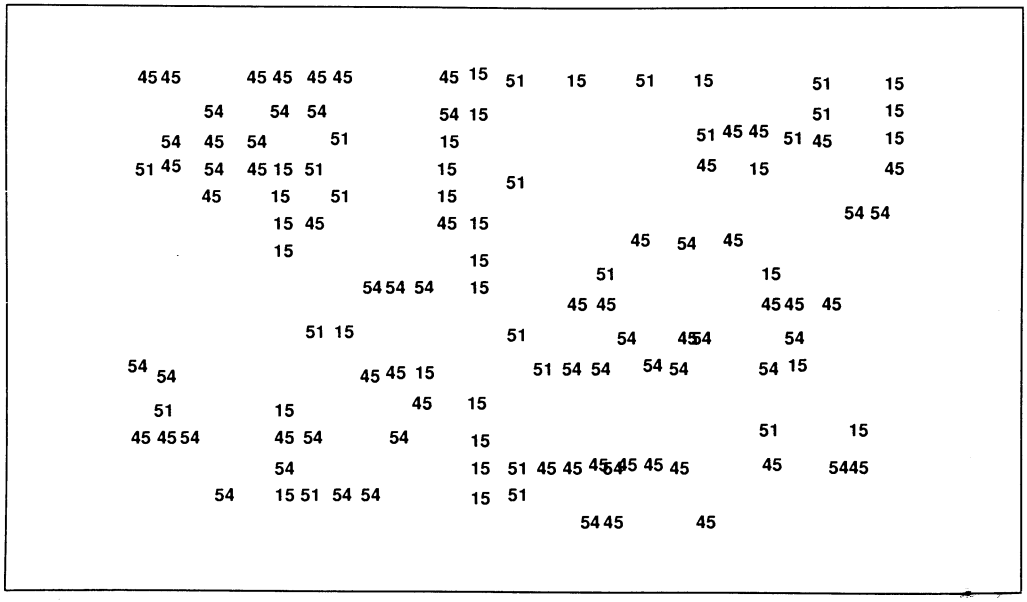

Fig. 6B. Spider values in same subarea of Census map

By looking only at "T" intersections in the area, Figures 7A and 7B, (and using knowledge that each "T" value corresponds to a single direction failing to "go through" – for instance "15" does not go through to the west or left), one may almost visualize the barriers (in this case known to be water inlets). A vertical string of "15's" just to the right of a vertical string of "51's" clearly flank one such inlet! A horizontal string of "54's" sitting above a similar string of "45's" clearly flank a horizontal inlet.

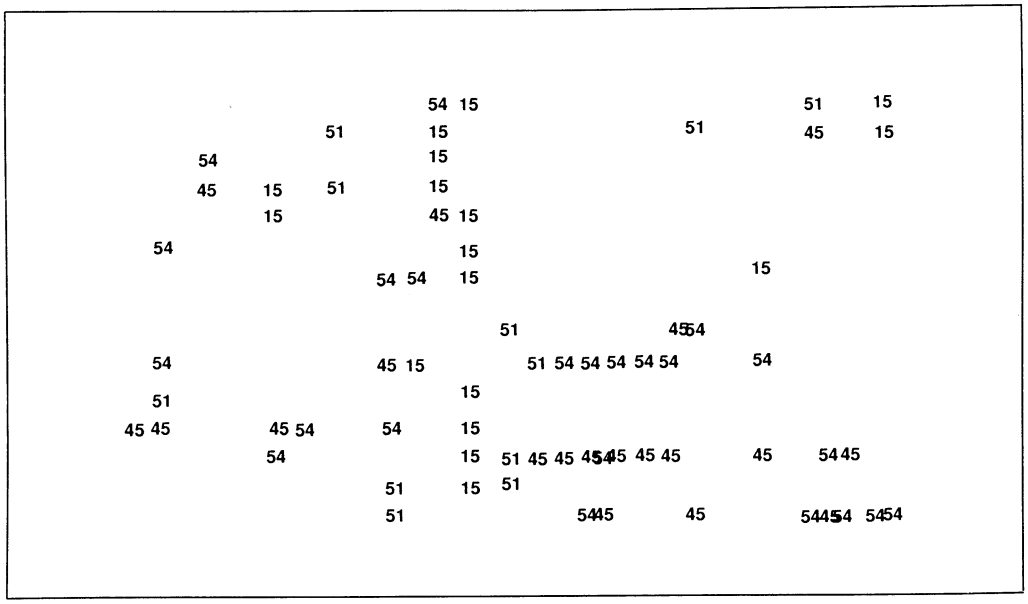

Fig. 7A. Intersections of USGS map with hexadecimal values {15, 51, 45, 54} (T's)

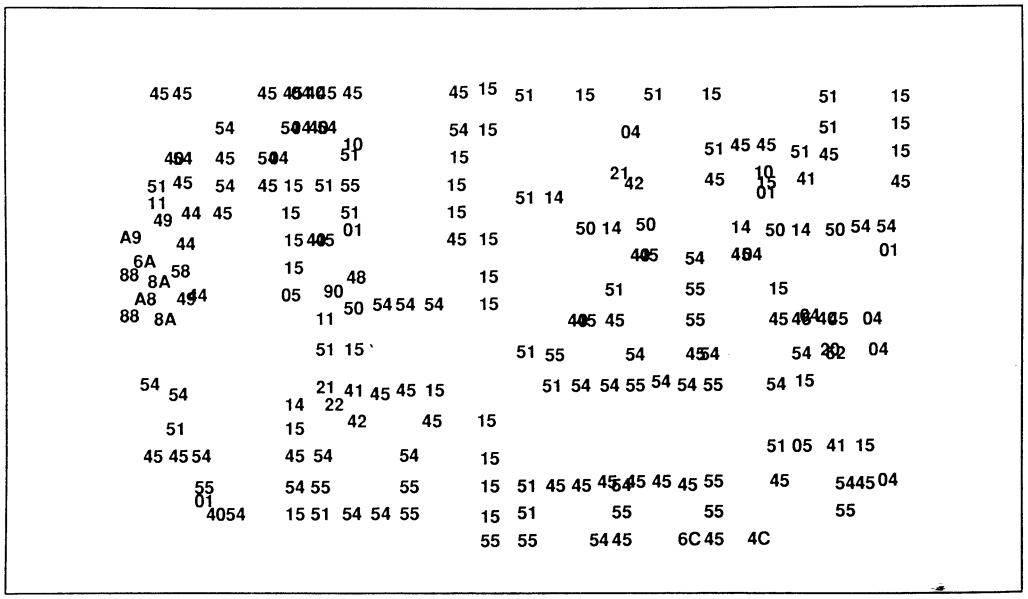

Fig. 7B. Intersections of Census map with hexadecimal values {15, 51, 45, 54} (T's)

A second filtering operation to reduce one's view to only a single class of intersections ("15's") produces a set of figures even more amenable to standard pattern recognition techniques.

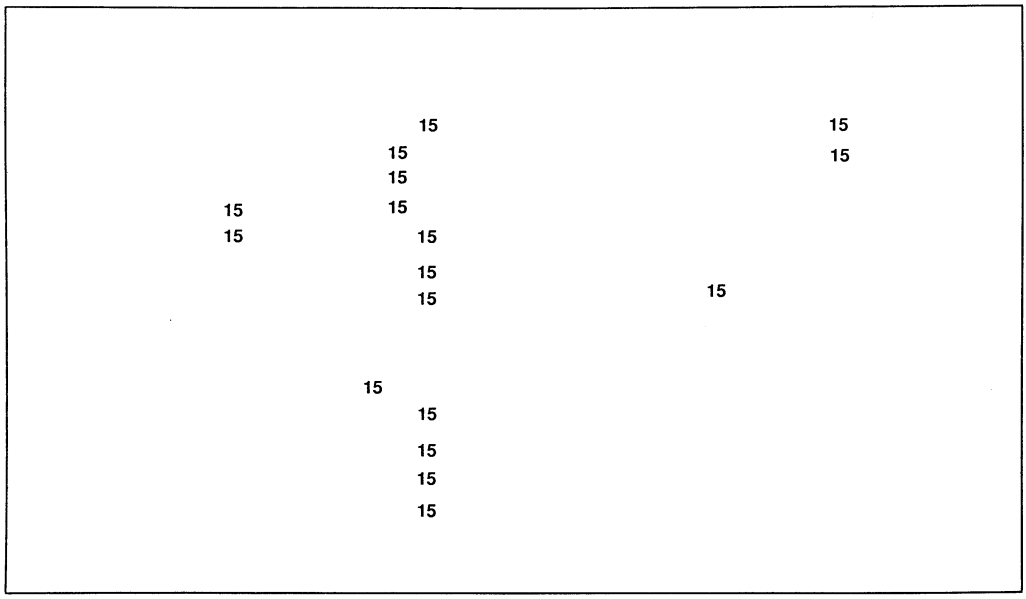

Fig. 8A. Intersections of USGS map with value = 15

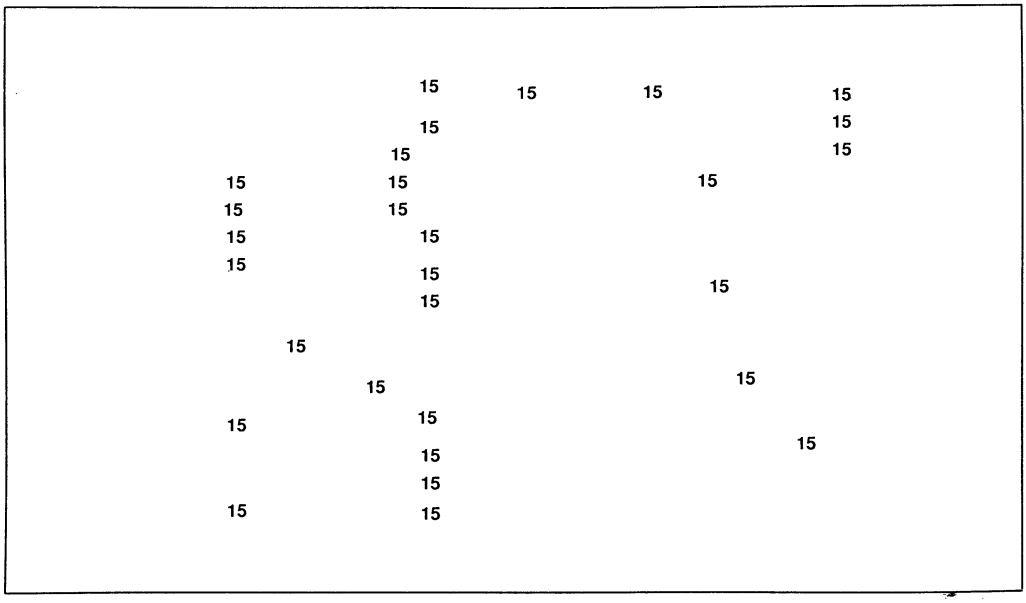

Fig. 8B. Intersections of Census map with value = 15

Although condensing the network information at an intersection to a single number inevitably causes some loss of information, the resulting patterns lend themselves to many standard pattern recognition and analysis techniques. The pattern distributions need to be viewed not only in terms of statistical error measurements, but also in terms of geometric relations of similarity and dependence shared by subsets of the spider function values. Two spider function values represent similar intersections patterns, for instance, if one value is twice the other or if their difference is a power of two. Likewise, values occurring at opposite ends of the same line segment must exhibit clear geometric dependence reflected in one of their digits. Only exploratory work has been undertaken to study geometric implications of spider function value distributions (Rosen and Saalfeld (1985)).

4. Conclusions

An analysis of distances between matching and nonmatching map features indicates that nearness measures can and should play a key role in automated map matching routines. A further link between computer cartography and spatial statistical analysis is provided by an integer-valued function defined on map intersection points. Preliminary exploratory work to study properties of this function has begun with limited data resources; and the approach used in that work has been outlined and illustrated here. The next stage in the research will involve the application of image analysis and pattern recognition techniques to attempt fully automated map matching.

5. References

Griffin, P. and White, M. (1985): Piecewise Linear Rubber-Sheet Map Transformations. The American Cartographer, 12(2), pp. 123–131.

Lupien A. and Moreland, W. (1987): A General Approach to Map Conflation. AUTOCARTO 8 Proceedings, Baltimore, MD, pp. 630–639.

Lynch, M.P. and Saalfeld, A. (1985): Conflation: Automated Map Compilation – A Video Game Approach. AUTOCARTO 7 Proceedings, Washington, D.C., pp. 342–352.

Rosen, B. and Saalfeld, A. (1985): Matching Criteria for Automatic Alignment. AUTOCARTO 7 Proceedings, Washington, D.C., pp. 456–462.

Saalfeld, A. (1985): Comparison and Consolidation of Digital Cartographic Databases Using Interactive Computer Graphics. Research Report Number 85–11, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

Received November 1986
Revised June 1988

Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 141–154

Data Collection With Hand-held Computers: Contributions to Questionnaire Design

Alois van Bastelaer,[1] Frans Kerssemakers,[1] and Dirk Sikkel[2]

Abstract: The newly designed Netherlands Labour Force Survey is conducted with hand-held computers on a continuous basis from January 1987. In March 1986 hand-held computers were tested in a pilot study; over 1 400 respondents from 700 households were interviewed. The test confirmed earlier findings that hand-held computers are accepted without any problems by interviewers as well as interviewees. Consistency checks were specified in some parts of the questionnaire. Inconsistencies had to be corrected by the interviewer. The quality of the questionnaire can be assessed by observing the interviewer's corrections and their paging backwards in the questionnaire (these manipulations were recorded by the computer). Inconsistencies remaining in the data when the interview was completed also suggest how the questionnaire can be improved.

Key words: Data editing; questionnaire design; CAPI; CATI; survey research.

1. Introduction

In March 1986 the Netherlands Central Bureau of Statistics conducted a pilot study to test a newly developed questionnaire for the Continuous Labour Force Survey with a hand-held computer (HHC). This experiment can be considered from two viewpoints. First, it was the logical continuation of two earlier experiments, described in Bemelmans-Spork and Sikkel (1985a, 1985b), where HHCs were tested in the Price Survey and the Consumer Expectations Survey. Second, the experiment was a preparation for the Continuous Labour Force Survey which started in 1987. This survey aims at measuring labour market flows. The monthly sample size is 10 000 addresses.

The questionnaire in the pilot study consisted of two distinct parts. The first part on household composition was more structured than the household section in common paper-and-pencil questionnaires. Because the "head of household" concept had to be

[1] Department for Statistics of Employment and Wages, Netherlands Central Bureau of Statistics, 6401 CZ Heerlen, The Netherlands.
[2] Research International Nederland B.Y.
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

Acknowledgment: The authors are grateful to the editor and the referees for their comments and suggestions which improved this paper.

avoided, a sequence of questions was designed to establish the relations between the household members. In the second part, all household members age 15 or older were interviewed about their current labour market statuses and (retrospectively) about their statuses in the past 12 months; no more than three jobs within the past 12 months could be described. These labour market histories are reconstructed by starting from the current labour market status and then going back in time. Many consistency checks were included in the questionnaire on household composition; consistency checks were not yet included for the labour market questions.

In this paper we describe some results of the pilot study, focusing on interviewer behaviour and questionnaire design. Section 2 sketches a short history of computer assisted data collection. In Section 3, the pilot study and the software for the data collection are described. Section 4 deals with the acceptance of the HHC by interviewers and respondents. In Section 5 the inconsistencies in the data are discussed and conclusions are drawn from an analysis of these inconsistencies. An example of the detection of an ill-defined concept from an analysis of inconsistencies is described in Section 6. Section 7 deals with paging backwards in the questionnaire. Section 8 concludes.

2. History

The history of computer assisted interviewing goes back to 1971, when AT&T sponsored a CATI (Computer Assisted Telephone Interviewing) survey to measure customer evaluations of telephone services (Nicholls and Groves (1986)). Since then, CATI has developed rapidly throughout the world. Today it is a common tool for commercial market research, governmental statistics, and scientific purposes. Despite its 17 year history, little research has been done on the methodology and data quality of CATI. In market research it seems that the main consideration for the introduction of CATI is cutting costs compared with face-to-face interviews, which usually implies that a CATI system must be flexible and very simple to use. Once such demands are fulfilled the users are satisfied and feel no need for further research. This experience was confirmed by Groves and Nicholls (1986), who in their comprehensive paper, stated that "... there is little reliable empirical evidence that CATI affects data quality. This absence is especially noteworthy in the context of the frequent expectation of data quality improvement of CATI."

CAPI (Computer Assisted Personal Interviewing, interviewing with HHCs) began its development when computers that were small and powerful enough appeared on the market. As observed by Shanks (1983) and Shanks and Tortora (1985), CAPI questionnaires have much in common with CATI questionnaires. CAPI and CATI are in essence a sequence of displays which depend on previous responses. This dependence may be complicated. In this way the construction of a questionnaire closely resembles the development of a computer program. House (1985) and Jabine (1985) observe that the design and documentation of a CATI questionnaire requires the same approach as ordinary computer software. A significant difference between CAPI and CATI is that CAPI does not impose extra limitations on the complexity of the questions and response categories and that CAPI allows for visual aids such as display cards.

The first test of HHCs that we are aware of was carried out by Statistics Sweden, in their Price Survey, using a pocket calculator sized computer (Danielsson and Maarstad (1982)). Later the Netherlands Central Bureau of Statistics started experimenting with HHCs. In Holland too the Price Survey was used for tests with a larger HHC, the size

of a telephone directory (Bemelmans-Spork and Sikkel (1985a)). Both experiments were successful in that they proved that the use of HHCs was possible, although the hardware needed improvement. From there on the strategies of Holland and Sweden diverged. Statistics Sweden formulated demands for an HHC that corresponded to its needs and found a manufacturer to develop this computer (Lyberg (1985)). The Netherlands Central Bureau of Statistics continued experiments with HHCs that had been developed for a more general market. Around Christmas 1984 a trial was carried out with a simple household survey, the Consumer Expectations Survey, see Bemelmans-Spork and Sikkel (1985b). In this experiment one group was interviewed with HHCs, and another with paper-and-pencil. Both groups consisted of about 175 respondents. The comparison showed no differences in unit nonresponse or item nonresponse for sensitive questions. Nor were there significant differences between the groups in the frequency distributions of the relevant variables. The first CAPI survey for production purposes was conducted in 1983 in a rather different context. Bus passengers in Durham, U.K., were asked simple questions to which the answers were entered into a computer that was a little larger than a pocket calculator. Hamilton (1985) reports that this survey was satisfactory with respect to the fieldwork as well as to economic aspects; the investments paid for themselves within a year.

3. The Pilot Study and the Software for Data Collection

Since the findings from the previous experiment were felt to be satisfactory, the pilot study of the Continuous Labour Force Survey was not designed to measure subtle differences between two different groups. Instead more quantitative evidence was sought about the acceptance of CAPI by interviewers and respondents and about any problems connected with the use of hand-held computers for data collection. A second objective was to obtain information on the data resulting from interviews with a complex questionnaire. The reduction of measurement errors was also analysed by comparing earlier false entries with the final correct entries in the same interview. In this way the pilot study was highly useful for the design of the final questionnaire to be used from 1987.

During the last three weeks of March 1986, 23 interviewers visited 1 224 addresses in ten municipalities. They had received a training of one full day and two half days. The interviewers used an HHC, type NEC PC-8201A with two memory banks of 32Kb RAM each. One bank contained the questionnaire and answers. The BASIC-program QUEST2 that interprets the questionnaire and a module for data communication were loaded on the other bank. QUEST handles the display of the question-texts and the response categories; it handles data entry and data storage in a compressed form which is to be decoded into fixed field records after the data are received by the central host computer. It checks the specified data inconsistencies and it manages the branching and skipping. It is the hardware, not the software that limits the number of questions. The definition of a question consists of four sections: in Section 1 the question type (precoded, numeric, open ended, etc.) was defined and the question text entered; Section 2 contained the response categories or the response range; Section 3 specified edit checks; and Section 4 contained the codes for branching and skipping. These codes can depend on the logical or numerical operations of previous entries. Text strings are often defined as variables for repeated use, also depending on previous answers. Besides saving memory, this allows

the clear-cut phrasing of questions and response categories. The detection of an inconsistency with prior answers causes the relevant questions to be displayed one after the other for confirmation or correction until the answers are reconciled.

The interviewers had a number of special programmable keys at their disposal in addition to the regular data-entry keys: return to the immediately preceding question (programmable key 1), return to any previous question (shift + programmable key 1), confirm a previous entry after having returned to a previous question (key 2), add remarks (shift + key 2), no choice in a multiple choice question (key 3), return to the current question while displaying all intermediate questions that have already been answered (shift + key 3), do not know (key 4), immediate return to the current question skipping intermediate questions (shift + key 4), refusal (key 5), refusal further cooperation with the interview (shift + key 5), consult table with household data (cursor right) and display question number and bytes still free (cursor left). Some programmable keys were confusing (e.g., return to the current question with and without displaying intermediate questions) and some were redundant (confirming an entry with key 2 whereas the enter/return key was used for the current question). In the revision of QUEST this user interface was redefined.

Each HHC was programmed to automatically phone the central host-computer at the office at a specified time during the night. These times were different for each HHC and distributed uniformly throughout the night. The data transmission took a few minutes, and the quality of the data transmission was examined through check sums. If the data transmission was successful, the data were released from the HHC so that the HHC could be used for new interviews. Following some problems during the first few days, the communication ran smoothly for the remainder of the test. This led to the conclusion that communication by phone works well (at least given the quality of the Dutch telephone network, which is fairly high). The value of tailor-made software for communication, however, should not be underestimated.

Due to the short fieldwork period, the total response rate was rather low: 56 %, i.e., 1 407 persons of age 15 or older participated.

4. Acceptance by Interviewers and Respondents

Danielsson and Maarstad (1982) and Bemelmans-Spork and Sikkel (1985a, 1985b) gave the impression that the HHCs were readily accepted by interviewers and informants. Due to the relatively large sample in our experiment, we now are able to confirm these impressions with more solid results from two evaluation forms filled in by the interviewers. One form was filled in for each responding household, the other after each week of interviewing.

The following questions were answered by the interviewer following each interview.

- Did the respondent's attitude change noticeably when you showed him or her the hand-held computer? (Table 1)
- Did the respondent comment on the use of the hand-held computer?
- If so, how?
- Did you feel that the hand-held computer caused any inconvenience for the respondent when answering the questions? (Table 2)
- Did the respondent inquire about the data processing or about confidentiality?

In addition to questions about the structure and content of the questionnaire the interviewers had to answer the following questions every week.

- Did typing texts for questions on economic activity or occupation interfere more with the interview than writing the answers on a paper-and-pencil questionnaire? Why?
- Did the hand-held computer refuse any answers (e.g., was the message "are you sure?" displayed)?
- Did you have to return to previous questions to correct mistakes?
- Were there any problems with returning to previous questions?
- Were there any problems with the hand-held computer?
- Could you enter answers that you already knew (e.g., on the compostion of a household) fast enough?
- Did the hand-held computer cause any delay when you wanted to ask a new question or enter the answers?
- Which questions caused problems?
- Were there any problems with the modem?
- Was the readability of the screen sufficient?
- Do you prefer working with a hand-held computer or a paper-and-pencil questionnaire? (Table 3)

First we shall give some results of the forms that were filled in per household. The interviewers were instructed to show the HHCs only after the respondent had agreed to the interview. In the evaluation form there was a question about the respondent's reaction, see Table 1.

Table 1. Respondents' first reaction to the HHC

	1986		1984	
	Absolute	%	Absolute	%
No reaction	667	92.4	113	65.3
Positive reaction (e.g., interested)	36	5.0	34	19.7
Neutral reaction (e.g., surprised)	9	1.2	21	12.1
Negative reaction (e.g., suspicious)	10	1.4	5	2.9
Total	722	100	173	100

Here the results are compared with the previous experiment in 1984. In neither was there any extra nonresponse due to the HHC. The negative reactions were almost negligible, and the following are typical examples: "automation strikes again" or "can we be recorded?" Examples of favourable reactions are: "very interesting," "called her husband because she thought it was fantastic," "how nice, is that a tape-recorder (or typewriter)?" and "will we be on television?" Most surprising, however, was the increasing number of respondents who did not react at all. This suggests that there is a growing acceptance of the computer as a common tool. A second question was: "Did you feel that the hand-held computer caused any inconvenience for the respondent when answering the questions?" The answers are displayed in Table 2.

Table 2. Did the HHC cause any inconvenience for the respondent?

	Absolute	%
No	701	97
A little	20	3
Very much so	1	0.1
Total	722	100

The vast majority of interviews presented no problems. Another question concerned confidentiality. About 100 respondents asked about this but most of their comments would also have been valid for paper-and-pencil interviews. They were the normal questions on the method of data processing, data protection, the retainment of anonymity, and the possibility or probability of linking interview responses to names and addresses. Other comments were more specifically related to the use of the HHC. Respondents expressed distrust of the way the answers were recorded, approval that the answers could not be changed once they had been entered in the HHC, and curiosity over how the data were transmitted to the office (a frequent question). Most respondents could be convinced that CAPI guaranteed confidentiality better than paper-and-pencil questionnaires because the answers are stored in a compressed form separate from the questions and because the answers are recoded, these new codes are not identical representations of the (alpha)numerical entries.

In the weekly evaluation forms, the interviewers were questioned about several aspects of CAPI. In the 1984 test it appeared that the quality of the light in the respondents' houses affected the readability of the HHC screens. In the current test there was a specific question about readability; about one in five interviewers complained about the poor readability of the screen.

In the first week, six interviewers felt that using the keyboard to enter text strings interfered more with the interview than writing answers on paper questionnaires. In the second and third weeks only two or three interviewers retained this opinion. The interviewers gradually became more accustomed to the keyboard for data entry.

During the first week almost half the interviewers complained about the slow speed of the program, especially when data had to be entered that did not need to be asked or were already known. This combined with complex skipping patterns and consistency checks slowed down the reponse time of the program to about two seconds. In the second and third weeks fewer interviewers complained about the speed; here too they grew familiar with the hand-held computer and the questionnaire. Meanwhile better hardware and other software (a Pascal instead of Basic program) have considerably improved the performance of an interview with the hand-held computer.

The hand-held computer did fail now and then, mostly because of lack of electrical power or because of program bugs or disconnected chips.

The reported problems may suggest that the interviewers had a bad attitude towards CAPI. This, however, was not the case as appears from a general comparison of hand-held computer and paper-and-pencil questionnaire (Table 3). Here we distinguish between the first and the third week of the experiment.

Table 3. Comparison of CAPI and paper-and-pencil

	Week 1		Week 3	
	Absolute	%	Absolute	%
CAPI better	11	52	10	71
Neutral	4	19	1	7
Paper-and-pencil better	6	29	3	21
Total	21	100	14	100

The percentage of interviewers who preferred CAPI to paper-and-pencil increased from 52 in the first week to 71 by the third week (the different totals of interviewers are caused by the fact that not every interviewer participated every week). Some favourable comments by the interviewers: "after two weeks better, after three weeks good, no longer uncertain," "much more convenient, no more paperwork," and one negative: "the interview is less natural."

5. Exploring the Questionnaire Design

5.1. Introduction

For each interviewer, the HHC also recorded important information about the flow of the interview, for example, returns to previous questions to consult the answers or to correct errors and answers prior to corrections. These data describe the error-checking and may point out ill-defined or poorly understood concepts in the questionnaire.

Error checks specify that the answer to question R must be in range Y if the answer to a previous question Q lies in range X. If such a condition is not satisfied, the HHC queries "are you sure?" and asks question R again. If the interviewer confirms question R, then the previous question Q is asked again. Only after confirming this question and once again confirming question R is the interviewer permitted to enter an inconsistency.

5.2. True value: the number of household members

A true value can be assessed if two different questions with a common content are asked and if the answer to the second question is redundant. The latter information can then be used as a check to the former answer and vice versa to determine the true value. A simple example from the household questionnaire illustrates how error checks can be defined on two questions with a common content. These questions are "how many household members are there?" and "is there anyone else in the household?" The latter question is asked repeatedly after the data for each person are completed. The answers must be consistent with the previously given number of household members.

Another error check consists of two questions about the number of household members and the household composition. A couple with children consists of at least three members, a couple with children and others of at least four. The purpose of the questions about the number of household members and the household composition is to give the interviewer a preliminary overview before she proceeds with the questions for each household member. The response categories

Table 4. Broad composition of the household

		Absolute	%
*	Single household	193	28
a.	(Un)married couple alone	180	26
b.	(Un)married couple + child(ren)	264	38
c.	(Un)married couple + child(ren) + other(s)	6	1
d.	(Un)married couple + other(s)	1	0.1
e.	Single parent + child(ren)	30	4
f.	Single parent + child(ren) + other(s)	2	0.3
g.	Other (household core not: (un)married couple or single parent)	10	2
Total number of households		686	100

and the corresponding frequency distribution of the second question are given in Table 4.

In this question the concept of the household core instead of the head of the household was central. In cases *a* to *d* this consisted of the couple, for *e* and *f* it was the single parent. The interviewers were instructed first to enter the data of the household core, then those of the children (of one or both members of the household core) and finally those of the others. The family relations were always a relative of a member of the household core, where possible the respondent.

The conflicts that actually occurred and the reactions of the respondent are displayed in Table 5. It appears that 46 conflicts were detected, 2 of which remained in the final data (confirmed twice). In one case the household was reported to consist of 4 members, after which the data of only 2 persons were entered. In the second case 33 members were reported followed by data on 3 persons. Altogether 36 conflicts involved the related questions of the number of household members and the question of whether there was still another household member. Of these, 33 were immediately resolved by changing the answer to the latter question.

Table 5. Conflicts with the number of household members

Current question	Conflicting answer to current question		Confirmation current question	Confirmation previous question	Total number of conflicts
Another person in household after	yes	no			
person 1	6	17	0	0	23
person 2	3	2	1	1	5
person 3	0	4	1	1	4
person 4	3	0	0	0	3
person 6	1	0	1	0	1
Composition household	Couple without children	Couple with children			
	9	1	0	0	10
Total			3	2	46

5.3. *True value: the household composition*

Another illustration applies to error checks on the household composition. If there are children in the household according to the preliminary household composition reported, then the third (for single parent families, the second) person should be a child. Actually the questionnaire was programmed in such a way that the interviewer had only to confirm that the next person was indeed a child. If this was denied, the conflict had to be solved. Also if there were no "others" then all persons not belonging to the household core must have been children. If categories *a* to *d* of the household composition (Table 4) were entered, then it could not be denied later without a conflict that there was a couple in the household, et cetera. The various categories of the household composition resulted in equally varied patterns of conflict, as described in Table 6. The question on the couple's marital status was asked

Table 6. Conflicts with the household composition

Current question	Conflicting answer to current question			Confirmation current question	Confirmation previous question	Total number of conflicts
	married couple	unmarried couple	no couple			
Unmarried or married couple	0	0	3	0	0	3
Child of person 1?	yes	no				
person 2	0	3		1	1	3
person 3	0	4		3	2	4
person 4	0	2		2	1	2
Relation to person 1	child	other				
person 5	2	0		1	0	2
Total				7	4	14

only when the household composition indicated that there was a couple. The answer "no couple," therefore, always created a conflict which in the pilot study was always solved. All remaining inconsistencies were between the reporting of children in the household composition and the not reporting of these children later on in the household box.

5.4. *Other conflicts*

The number of other conflicts was so small that we do not present them in a table. These conflicts concern the network of relations established within the household box. In five cases it appeared that the relation of person B to person A was in conflict with the marital status of person A. This relation was corrected three times; the marital status was corrected once; one inconsistency was confirmed. This concerned a remarried widower who still wanted to be considered a widower. Finally two conflicts appeared between the marital status of person B and the question about a married or unmarried couple. In both cases the marital status of person B was changed to "married."

5.5. *True value: the date of an event*

The software can handle a variety of labour market histories with a maximum of three jobs in the past twelve months. One respondent may have had a single job for a few years already, another respondent may have changed jobs twice in the past twelve months while being unemployed between two jobs. These event histories were established in the questionnaire by introductory questions which determined the correct path through the questionnaire. The response categories of the introductory questions were three or six month periods. After the introductory questions, the specific dates of beginnings and terminations of jobs were asked. Obviously, these dates have to satisfy certain order relations. A new job should have started after a previous job started, et cetera. However, no error checks were specified. In the absence of error checks some inconsistencies remained (Table 7).

The inconsistencies of cases 3, 4, and 10 are violations of the order relations of the dates. They are caused mainly by respondents who usually do temporary work and therefore have a complicated labour market

Table 7. Inconsistencies of dates with the introductory questions and violations of chronology

	General labour market history	Inconsistencies and violations	Specific labour market history
1.	B	start current job less than 1 year ago (4x)	C
2.	C	start current job more than 1 year ago (2x)	B
3.	C	start current job after date of interview (1x)	
4.	D	start last job coincides with end last job (5x)	
5.	D	end last job precedes start last job (1x)	
6.	F	start previous job more than 1 year ago (1x)	G
7.	G	end previous job coinsides with start current job (1x)	
8.	H	start previous job more than 1 year ago (1x)	I
9.	J	start earliest job more than 1 year ago (1x)	K
10.	K	start previous job coincides with start current job (1x)	

For the meaning of the capitals indicating the type of history, see the Appendix.

history. Case 7 is not a violation of an order relation, but rather a false entry. The end of a previous job coincides with the start of a current job even though data are also provided on the intermediate period.

The inconsistencies of cases 1, 2, 6, 8, and 9 are of another type. Here the broad indication does not correspond with the specific dates.

Table 7 may give the impression that the number of inconsistencies is insignificant. This is true if compared with the total sample of 1 407 respondents. However, if there are, say, 6 inconsistent records among the 19 records with a single job in the last 12 months that started and ended within these 12 months (labour market history D), then error checks are necessary for more complex labour market histories.

5.6. *Interpretation*

In CAPI and CATI, consistency checks have two purposes. First, they serve to ensure that the answers are entered correctly. If any errors are detected by the HHC program, they can be rectified during the interview. Consequently, the resulting data set is error free and may even be statistically processed directly. Second, with computer assisted interviewing, data consistency is essential because responses or series of responses lead to intricate branching and skipping patterns later in the interview. Interviewers are expected to perform better because the hand-held computer takes care of the routing of the questions and because questions are displayed one at a time, thus focusing the interviewers' attention on that particular question. From time to time, however, inconsistencies are inevitable. Our most striking example was the married man who insisted on being registered as a widower. The HHC must allow for such a conflict, but it makes the task of the questionnaire designer far from easy.

A study of the conflicts that occurred in the field can contribute to developing the methodology of questionnaire design. In this way the designer may develop an intuition about the effects of error checks. The results of our pilot study are insufficient to draw definite conclusions, but they are sufficiently suggestive to generate a few recommendations. These recommendations may be generally valid, i.e., independent of the interviewing mode.

If error checks are specified between two different questions with common content, a true value can be assessed. When a true value is assessed during the interview, the measurement error is reduced. The pilot study of the Labour Force Survey provides evidence about the assessment of a true value. Because error checks can be specificed between two questions with common content, both questions should be retained in the questionnaire. Error checks are also effective when the answers to general and detailed questions have to be reconciled.

The number of corrections per interview may be interpreted as an indicator of the quality of a questionnaire. However, questionnaires on different topics cannot be compared since some topics are simply more complex than others. Hence a questionnaire – *ceteris paribus* – with the least number of corrections needed, is preferred.

If a survey organization wants to test a questionnaire, the software for computer assisted interviewing should be able to keep a record of all interviewer actions or, even better, to produce summary statistics on consistency checks.

6. Detecting Ill-defined Concepts

The occurrence of inconsistencies sometimes indicates an ill-defined or poorly understood concept. A surprising and important example is the parent-child relation in our definition of household. Many conflicts and most remaining inconsistencies had to do with this relation.

The concept "child" may have more than one interpretation in an interview and should therefore be well-defined. There are different reasons for confusion. The word "child" may be interpreted as the opposite of "grown up" or "married child." This may also create confusion in determining the household core (e.g., a single parent household core implies the presence of a child). Moreover there is the problem of adopted children or stepchildren, e.g., whether for a married couple a child of the second partner from a former marriage and thus of the household core according to the first question, will also be considered a child of the first partner to whom the relation is determined in the second question (by confirmation).

7. Detecting Deficiencies in a Questionnaire: Paging Backwards

A difference between computer assisted interviewing and paper-and-pencil interviewing is that the interviewer loses her overview to some extent because she can see only one screen at a time. With the HHC thr interviewer could page backwards, question by question, using a programmable key. This option not only gives the interviewer an overview when necessary, but is also used to correct previous answers. In this way, paging backwards can be considered a form of error checking that was not specified by the designer of the questionnaire. The HHC kept a record of the interviewer's backwards steps in the questionnaire.

Table 8. Number of steps needed to reach the desired question

Steps	Number	Percentage	Cumulative
1	664	85.46	85.46
2	67	8.62	94.08
3	21	2.70	96.79
4	8	1.03	97.82
5	6	0.77	98.59
6	1	0.13	98.72
8	4	0.51	99.23
9	2	0.26	99.49
11	1	0.13	99.62
12	1	0.13	99.74
17	1	0.13	99.87
29	1	0.13	100.00
Total	777	100.00	

Table 9. Questions that were referred back to more than 10 times

Question no.	Short description	Number of times back to	Number of times asked	Percent back
1	Result household: response/nonresponse?	12	686	1.75
2	Number of members of household?	14	686	2.05
107	Which member of the household answers questions?	17	1400	1.21
109	Activities of respondent (multiple choice)?	44	1407	3.13
111	Do you have a paid job?	12	787	1.52
112	Have you ever had a paid job?	11	733	1.50
121	Name of the company where you work?	15	598	2.51
122	Address of the company where you work?	18	598	3.01
124	In which department or place of the company do you work?	16	598	2.68
169	What is your occupation?	10	598	1.67
170	What are your main activities?	13	598	2.17
179	How many hours do you work (SWW[1])?	27	281	9.61
180	How many hours do you work (no SWW[1])?	18	306	5.88
181	Did you work longer last week?	22	598	3.68
184	Full time or part time?	11	598	1.50
213	Did you have another job before this one?	11	656	1.95
343	How long have you been looking for this (2nd) job?	10	77	12.99
372	How long have you been looking for a job?	13	60	21.67
423	Have you been looking for a job during the last months?	13	630	2.06
445	Sequence number address?	10	686	1.46
448	Number of households at this address?	14	686	2.05
459	Day of first visit?	18	686	2.62
460	Time of first visit?	14	686	2.05

[1] SWW = shorter working week.

The average number of times per household an interviewer paged backwards was 1.132. Table 8 shows how many steps the interviewers had to take to reach the desired question. In most cases one step was sufficient. In 67% of the cases paging backwards served to correct a previously given answer, and in 33% of the cases the previous answer was confirmed. This is consistent with findings in the CATI case reported in Groves and Nicholls (1986).

Since the "previous question" key could be used anywhere in the questionnaire, it is impossible to list all the 172 questions that were paged back to without reproducing a substantial part of the questionnaire. We therefore restrict ourselves to those questions to which the interviewers jumped back 10 times or more. In Table 9 these questions are described briefly. An indication is also given of how many times the question was asked (obviously you cannot go back to a question that was not asked). Moreover, the percentage of "back jumps" is indicated. This leads to some remarkable results. The question that was most frequently jumped back to (109) has a relatively low "back jump" percentage because this question was asked to everyone. The percentage is much higher for questions 343 and 372 because only few respondents abswered these questions.

A first conclusion of the analysis of paging backwards may be that the interviewers are concerned with the quality of their work. This is apparent from their frequent use of the "previous question" key. It was most often used to correct previously given answers. Given the preliminary status of the

questionnaire we shall restrict ourselves to only a few topics.

In question 109, the respondent was asked to report his activities in a multiple choice question (paid work, student, housewife et cetera). In question 110 the respondent had to make a single choice from the same alternatives (the most important activity). The fact that the interviewer paged back 44 times shows that this construction was poorly understood. In the questions determining the respondent's occupation the interviewers often paged back, indicating that this is a difficult subject. But on the other hand it is hard to improve such questions. Questions about working hours also cause problems especially when there is a shorter working week (questions 179–181). The retrospective questions 343 and 372 have the highest percentage of referrals to the "previous question." Apparently it is difficult to answer questions about a complicated labour market history.

The relatively low number of referrals to the "previous questions" in the household box is remarkable, but it may be explained by the relative simplicity of the subject. An alternative explanation is the abundance of consistency checks in the household box. Detected inconsistencies mostly lead to correction (see Section 5), so that there will be less need for correction by paging back.

Consistency checking and paging back are probably not independent processes. Also, when certain questions are spontaneously corrected by the interviewer, these questions may be identified as error prone, or, from another perspective, questions where error checks are effective.

It may also be asked whether the number of times the interviewers paged back is a good indicator of the quality of the questionnaire. As with error checking, the answer is not a simple "yes." The frequency of paging back is probably a good indicator of the difficulty of a questionnaire. Of course it is important to keep questionnaires as simple as possible, but there are simple topics as well as difficult topics, such as occupation or economical activity, about which we have to gather information. Therefore, the preferred questionnaire is one in which the interviewers page back least often, given the complexity of the subject.

8. Conclusion

This paper had two goals. First, it aimed at presenting new findings about acceptance and appreciation of CAPI by interviewers and respondents. Acceptance by respondents is no problem; to them CAPI circumvents rather than arouses suspicion about confidentiality. Interviewers' attitudes are also favourable: the majority of the interviewers consider a hand-held computer more convenient than paper-and-pencil.

The second goal was to contribute to the development of a methodology for CAPI. The results suggest that error checks in CAPI are necessary. This is consistent with Tortora (1985) who compared the use of error checks for CATI and paper-and-pencil. Moreover, keeping a record of the interviewer's consistency checks and movement through the questionnaire is one step in evaluating the quality of the questionnaire.

9. References

Bemelmans-Spork, M. and Sikkel, D. (1985a): Observation of Prices With Hand-held Computers. Statistical Journal of the United Nations Economic Commission for Europe, 3, (2).

Bemelman-Spork, M. and Sikkel, D. (1985b): Data Collection With Hand-held Computers. Proceedings of the 45th session. International Statistical Institute, Book III, topic 18.3.

Appendix. The types of labour market histories and their distribution; having a job is indicated by x; not having a job by –.

Labour market history	n		1 year before interview	Date of interview
A	682		———	———
B	536		xxxxxxx	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
C	63		———	———xxxxxxxxxxxxxxxx
D	19		———	———xxxxxxxxxxxxxxxx———
E	44		–xxxxxxxxxxxxxxx———	
F	6	[1][1]	———	———xxxxxxxx———xxxxxxxxxxxx
G	46	[35]	–xxxxxx	xxxxxxxx———xxxxxxxxxxxx
H	2	[1]	———	———xxxxxxxxx——xxxxxxxxx———
I	3	[2]	–xxxxxx	xxxxxxxxxxxxx———xxxxxxxxx———
J	1	[1]	———	——xxxxxxxx——xxxxxx——xxxxxxxxxx
K	4	[2]	–xxxxxx	xxxxxxxx———xxxxxx——xxxxxxxxxx
L	1	[1]	———	——xxxxxxxx——xxxxxx——xxxxxx——

[1] Two jobs may succeed each other immediately from labour market history F; this distinction is not made in the figure, i.e, a period indicated by "—" may be empty. The number in [..] indicates the number of persons with immediately succeeding jobs.

Danielsson, L. and Maarstad, P.A. (1982): Statistical Data Collection With Hand-held Computers – A Test in Consumer Price Index. Statistics Sweden, Örebro.

Groves, R.M. and Nicholls, W.L. (1986): The Status of Computer-assisted Telephone Interviewing: Part 2 – Data Quality Issues. Journal of Official Statistics, 2 (2), pp. 117–134.

Hamilton, T.D. (1985): Hand-held Data Capture Devices. Paper presented at the meeting of the Study Group on Computers in Survey Analysis, London.

House, C.H. (1985): Questionnaire Design With Computer Assisted Telephone Interviewing. Journal of Official Statistics, 1 (2), pp. 209–220.

Jabine, T.B. (1985): Flow Charts – A Tool for Developing and Understanding Survey Questionnaires. Journal of Official Statistics, 1 (2), pp. 189–208.

Lyberg, L. (1985): Plans for Computer Assisted Data Collection at Statistics Sweden. Proceedings of the 45th session. International Statistical Institute, Book III, topic 18.2.

Nicholls, W.L. and Groves, R.M. (1986): The Status of Computer-assisted Telephone Interviewing: Part 1 – Introduction and Impact on Cost and Timeliness of Survey Data. Journal of Official Statistics, 2 (2). pp. 93–115.

Shanks, J.M. (1983): The Current Status of Computer-assisted Telephone Interviewing: Recent Progress and Future Prospects. Sociological Methods & Research, 12 (2), pp. 119–142.

Shanks, J.M. and Tortora, R. (1985): Beyond CATI: Generalized and Distributed Systems for Computer assisted Surveys. Proceedings of the first annual research conference of the Bureau of the Census, U.S. Bureau of the Census, Washington D.C., pp. 358–377.

Tortora, R. (1985): CATI in an Agricultural Statistical Agency. Journal of Official Statistics, 1 (3), pp. 301–314.

Received December 1986
Revised March 1988

Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 155–172

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Self/Proxy Response Status and Survey Response Quality

A Review of the Literature

Jeffrey C. Moore[1]

Abstract: Three decades of research have not produced conclusive evidence of consistent response bias or response error variance differences due to self/proxy response status. The net nonresponse effect may also be close to zero due to compensating effects for the various components of nonresponse. The main cause of the lack of evidence is the methodological shortcomings of much of the research which purports to address the self/proxy issue. In addition, the few methodologically sound studies – most importantly, those which control potential self-selection biases, and whose subject matter makes the self/proxy distinction appropriate – in general have produced no effects or conflicting effects (or, in the case of nonresponse, compensating effects). However, lack of convincing evidence of quality differences is not synonymous with convincing evidence of no quality differences. Until more data are gathered, the conclusion that self and proxy survey responses are of equivalent quality must remain tentative.

Key words: Self/proxy response; response quality; response bias; response error variance; nonresponse; survey design.

1. Introduction

Survey research involves many compromises, of which sampling is perhaps the most fundamental. Sampling forces the survey designer to accept reduced estimate precision in exchange for cost and effort feasibility. Other compromises are more subtle. For example, clustered sample designs sacrifice some of the information value of individual responses for enhanced data collection effi-

[1] Center for Survey Methods Research, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A. The views expressed herein are the author's and do not necessarily represent official views or policy of the Census Bureau. I thank Kent Marquis, Dan Hill, and Betsy Martin for their many helpful and encouraging comments on early drafts of this paper. The final paper has also benefited from the comments of three anonymous *JOS* reviewers.

ciency. Questionnaire design decisions must balance the survey designer's desire for complete information against a reasonable respondent burden. In retrospective surveys, decisions about the length of the reference period attempt to balance the advantages of a short reference period (presumably, reduced memory errors) against a longer period in which the naturally greater number of target events serves to improve estimate precision. Follow-up efforts to reduce nonresponse push against budget and schedule constraints.

In addition to the design decisions they share with all sample surveys, surveys which collect data on each eligible member of each sampled household require one more key design decision: how much effort to expend gathering data *on* individuals *from* each individual himself or herself. The solution to this problem has always varied from survey to survey, even within the same survey organization. For example, the Census Bureau's four major continuing household sample surveys – the Current Population Survey (CPS), the National Health Interview Survey (NHIS), the National Crime Survey (NCS), and the Survey of Income and Program Participation (SIPP) – use four very different sets of procedures to collect data on each household member. As the costs of sample surveys have increased, however, there seems to be a trend to rules more accepting of proxy responding. And yet, despite at least three decades of concern about the effects of self/proxy response status on data quality, survey designers have little evidence to justify the use of more costly procedures or to defend the quality of data obtained less expensively.

The purpose of this paper is to review the literature for evidence on the data quality effects of self versus proxy response status. The central question can be stated in general terms as follows:

> For survey items about individual A, are there systematic differences in quality between the responses obtained directly from A versus those obtained from some other respondent who is reasonably likely to be informed?

The common wisdom about self/proxy effects is that the best information about person A will come from person A directly (Sudman and Bradburn (1974); Roshwalb (1982); Mathiowetz and Groves (1983)). There are, of course, recognized exceptions. Proxy responding is generally permitted – if not required – for children and for those too mentally or physically infirm to respond. But this review is concerned with the "standard" proxy situation, in which an eligible and capable adult for some reason does not self-respond. The generally accepted notion is that response quality suffers to the extent that such persons do not respond for themselves. In its common form, then, the core question of this review is explicitly directional:

> Do survey data suffer in quality when eligible sampled persons do not respond for themselves? To what extent is quality sacrificed when a survey designer opts for rules which permit proxy response?

2. Methodological Considerations

The assumption of quality differences favoring self-response has intuitive appeal, although in certain instances the opposite case can be made, such as when the survey subject matter may evoke self-presentation pressures (e.g., Berk, Horgan, and Meyers (1982)). However, the standard assumption also seems to draw support from the self/proxy literature. Unfortunately, a substantial portion of this literature only appears to address the data quality implications of alter-

native respondent rules; close inspection often reveals conceptual and methodological shortcomings which render conclusive judgments about self/proxy quality effects impossible. Much of the research which purports to address the question of self/proxy response quality differences falls short on at least one of three important criteria: (1) survey subject matter, (2) self/proxy status control, and (3) quality assessment.

2.1. Survey Subject Matter

The most obvious criterion for the examination of self/proxy effects is that the survey inquiry must refer unambiguously to an individual. If this criterion is not satisfied, the self/proxy distinction is not meaningful. There is an extensive "pseudo proxy" literature which probably contributes to the consensus judgment regarding self/proxy effects. For example, husbands and wives have been found to provide very different accounts of their relative influence in household purchase decisions (Ferber (1955a)), and of family economic characteristics in general (Ferber (1955b)). Spouses' reports of frequency of intercourse (Levinger (1966)) and of other "shared experiences" (Mudd, Stein, and Mitchell (1961)) show large discrepancies. Parents and children disagree about past childrearing practices (Radke (1946); Kohn and Carroll (1960)).

There are two main problems with this literature. First, although respondent discrepancies seem to be the general rule, the data are by no means uniform; many studies have found substantial agreement between respondents. (See for example, Neter and Waksberg (1965) on household expenditures; Haberman and Elinson (1967) on family income; Kinsey, Pomeroy, and Martin (1948) and Rutter and Brown (1966) on sexual behavior and other relationship variables.)

The second and more critical problem is that the use of the terms "self" and "proxy" is simply not appropriate for these topics. Many survey questions are about the past behaviors, life events, and current circumstances and characteristics of people not as individuals but as participants in some collectivity – spouses, families, households. While reporting consistencies and discrepancies among collectivity members pose interesting questions for survey methodologists, it is not legitimate in these investigations to identify any particular member as a "self" respondent. The literature searches carried out for this review captured studies of this type with some frequency, suggesting that conventional wisdom about self/proxy differences may be contaminated by irrelevant research findings.

2.2. Self proxy treatment control

The second criterion concerns proper research design. Unfortunately, a common "design" in self/proxy research is no design at all – a survey is conducted, some people respond for themselves and others are responded for by proxy, the responses of the two naturally-occurring groups are compared, and conclusions are drawn about the effects of response status on response quality. Without strong assumptions, however, such conclusions are not justified.

Definitive research evidence can only derive from studies which can dispense with reasonable competing explanations for observed effects. Studies of naturally-occurring self/proxy effects are open to the possibility that observed differences are a result of self-selection biases and do not necessarily indicate response quality problems for one or the other group. Thus, the typical finding in health surveys of more frequent reporting of health conditions, doctor visits, hospital stays,

etc. for self-responders than for those responded for by proxy (e.g., Horvitz (1952); Enterline and Capt (1959); Linder (1959) (cited in Cartwright (1963)); Haase and Wilson (1972); Berk et al. (1982)) may simply reflect the greater likelihood of finding the less healthy household members at home when an interviewer calls. (Haase and Wilson (1972), Kovar and Wright (1973), Berk et al. (1982), and others who have identified this effect have noted the possibility that true sample differences account for the reporting differences of self and proxy respondents.)

Although procedures to assess response quality (see Section 2.3.) make tantalizing additions to uncontrolled treatment studies, they do not render the evidence on response status effects any more conclusive. Such studies still cannot discount self-selection sample bias as a possible explanation for observed quality differences. Thus, studies showing more complete survey reporting of medical conditions for self-respondents than for proxies as judged against a subsequent medical examination (Commission on Chronic Illness (1957); Elinson and Trussell (1957); Krueger (1957)), for example, or more accurate income reporting according to a match to administrative records (Kilss and Alvey (1976)), are fundamentally uninformative because they fail to address the key question: if the original proxy group had responded for themselves, would their reports have shown any greater correspondence to the validating data?

Of course, this difficulty afflicts all uncontrolled self/proxy studies, regardless of the direction of their results. Thus, the conclusion of Berk et al. (1982) that proxies produce *better* quality reports of stigmatizing physical conditions also goes beyond the available data. Unless self/proxy status is controlled, the conclusion that quality suffers – or is enhanced – as a result of proxy status rests on an untested assumption of self and proxy sample comparability.

2.3. *Response quality assessment*

The third methodological criterion is a concrete assessment of data quality. A long-standing tradition in survey nonsampling error research is that a definitive evaluation of response quality cannot occur without reference to a "true value"(Hansen, Hurwitz, Marks, and Mauldin (1961)), or "some external criterion" (Sudman and Bradburn (1974)). This approach to data quality, then, is concerned with the deviation of individual survey responses from some external standard of truth.[2] Aggregated across a set of responses, these deviations can be used to assess the total error associated with self and proxy response, and the extent to which response errors under the two conditions are systematic or random.

Response quality is best evaluated through a comparison of individual survey responses with some independent, external criterion, such as existing records, or an objective, non-survey-based measurement of the same phenomenon. Since a well-designed validity assessment is difficult to execute, many self/proxy studies have taken the easier course of simply comparing the aggregate responses of the two respondent groups. Such studies typically adopt a "more-is-better" assumption (or the reverse for socially desirable subject matter), and occasionally even a "self-is-true-therefore-proxy-differences-indicate-error" assumption. The latter simply assumes what should be a matter for objective inquiry. The former probably derives from the results of one-directional or partial record check designs, in which only a limited range of survey reports is validated. As Marquis (1983) and Marquis, Duan, Marquis,

[2] An important corollary of this definition of quality is that the subject matter of eligible studies must offer at least the potential of external verification, which excludes survey measures of "subjective phenomena" (Turner and Martin (1984)).

and Polich (1981) point out, such designs are guaranteed to produce apparent bias estimates in only one direction.

Studies lacking a direct assessment of response quality can still yield useful information about self/proxy effects. If controlled response status treatments produce reliable self/proxy report differences, then we may infer that a stable response quality effect exists, even though the critical details of that effect (What are the directions and magnitudes of the biases? Which group's data are better?) must await additional research. However, a finding of no consistent self/proxy report differences does not necessarily indicate quality equivalence. The systematic "noise" associated with each response status may be comparable, but the random noise may differ greatly.

Because quality assessment is difficult to implement (and sometimes seemingly impossible), researchers with otherwise appropriately designed studies have often fallen back on more indirect quality indicators. The most frequent of these are the various dimensions of response completeness – item nonresponse, person nonresponse, and household nonresponse. Studies which consider the effects of self/proxy status on these (and other) indirect quality indicators are also included in the review.

3. Literature Review

Survey procedures which attempt to maximize self-response are more expensive than those which are more tolerant of proxy responding, and the difference can be enormous for large survey programs. By not requiring self-response, total CPS costs are effectively reduced by about 12 percent (Deighton (1967)) – or $150 000 per month at current rates. Estimates for other surveys have indicated that maximum self-response rules would increase costs by 5 percent for a comprehensive income survey (Kulka (1982)), 17 percent for a health survey (Kovar and Wright (1973)), and up to 30 percent for a Canadian labor force survey (Singh and Tessier (1975)).

These studies leave little doubt that self-response rules are the more expensive option. Do they buy better quality data? The remainder of this paper reviews the evidence on self/proxy response quality differences for three dimensions of response quality – response bias, response error variance, and nonresponse. In keeping with the discussion in Section 2, I consider here only those studies which meet the appropriate subject matter and self/proxy control criteria. Although quality assessment procedures are desirable, otherwise well-designed studies lacking this component are still informative, and so are also included in this review.

3.1. Response bias effects

3.1.1. Report level differences

A large portion of the self/proxy literature consists of studies which control the assignment of respondents to self or proxy response but which lack a direct assessment of response quality. These studies offer little evidence of consistent differences in overall report levels attributable to response status, and thus suggest no major differences in the extent of systematic error.

Crime surveys are a possible exception to this general rule. Turner (1972) presents results of the only known controlled self/proxy study in this area. Each household in a large sample (n=10 000) was randomly assigned to either a household respondent or a self-response treatment. For all eight crimes examined – strong-arm robbery, armed robbery, robbery attempts, aggravated assault,

simple assault, assault attempts, rape, and rape attempts – the self-response treatment produced more incident reports. Although Turner presents no significance tests, the differences are substantial. Again, these results only indicate that the biases associated with self and proxy reporting of criminal victimization are different, and not that one status elicits better data than the other.[3]

Similar investigations have been carried out in labor force surveys, although here the evidence suggests no response bias differences. An experiment comparing self-response to the standard CPS household respondent procedure produced very small (and probably nonsignificant) differences in labor force participation and unemployment rates (Deighton (1967)). Williams (1969) and Jones and Aquilino (1970) report results of another experiment in the CPS, involving dual interviews with two designated respondents, each of whom reported for self and all other household members. The experiment yielded near perfect agreement between self-reporters and proxies on labor force participation and unemployment rates.

A more complex CPS experiment was conducted a decade later, including a comparison of the standard household respondent procedure with a designated household respondent and a maximum self-response procedure. There were scattered significant interactions involving the respondent treatment variable. However, keeping the other experimental factors constant, there were no significant response status differences in estimated unemployment rates (Cowan, Roman, Wolter, and Woltman (1979); Roman and Woltman (1980)).

Martin and Butcher (1982) present results of a self/proxy experiment involving U. K,'s Labour Force and National Dwelling and Housing Surveys, in which independent interviews were conducted with both proxies (usually wives) and self-respondents about the latter's labor force activity and related issues. Martin and Butcher report very high levels of agreement between self and proxy reports across a wide range of topics, including labor force status, type of occupation, transportation to work, hours worked last week, age, completed education, etc.

Self/proxy research has a long history in the health survey area. In perhaps the earliest experimental examination of self/proxy treatment differences, Cartwright (1957), in a small pilot study, compared husbands' self-response health reports against those of their wives acting as household respondents. The average number of reported illnesses per husband was three times greater under the self-response treatment than under the proxy treatment (2.5 versus 0.8), and the proportion reporting no illnesses was about one-sixth as great (8% versus 49%).

Most early health studies, however, report no self/proxy report differences or inconsistent differences. Enterline and Capt (1959) randomly assigned adult males not at home on an initial visit to an immediate proxy interview or a self-response followup. They found no significant treatment differences for any of the ten specific chronic conditions examined, and the average number of conditions reported per person was identical for the two procedures. Items assessing parents' histories of heart disease also showed no consistent differences. Proxy estimates of daily fat intake were marginally higher than those obtained via self-response, but items concerning age, weight, stature, weight gain, and

[3] The quality implications of differing report levels are particularly ambiguous in the criminal victimization area. On the one hand, the common assumption is that victimization is generally underreported (thus, more reporting means better reporting). On the other hand, there is also great concern about "forward telescoping," in which incidents which occurred prior to the stated reference period are reported as having occurred within the reference period (thus, more reporting may mean worse reporting).

recent medical care showed no significant differences.

Similarly, in an early experimental pretest of the NHIS, adults in a sample of households were randomly assigned to either a strict self-response treatment, or a treatment which accepted proxy responding (Nisselson and Woolsey (1959)). The overall illness rate was significantly greater for the self-response treatment than for the proxy treatment, although the results are quite inconsistent across sex, age, and illness categories. Reported days of disability were consistently lower in the self-response group, although with only scattered significant effects. The authors' main conclusion, however, is that "the sampling and response variability in the data are too high to permit definitive conclusions as to possible biases in the use of a household respondent" (Nisselson and Woolsey (1959, p. 72)).

Kovar and Wright (1973) carried out a similar experiment, on a much larger scale, comparing a maximum self-response treatment with a more lenient "accept-proxy" treatment. For most items there appear to be no significant treatment differences,[4] although reporting is fairly consistently higher under self-response conditions. Only two items – "limitation of activity" reports, and reports of recent doctor visits – show a significant treatment effect, with the self-response treatment producing significantly higher report levels. Although "limitation of activity" is clearly not identical to Nisselson and Woolsey's (1959 op. cit.) "days of disability," there does appear to be some inconsistency in these two sets of results.

[4] The authors contend that three of the two-week recall items show significantly higher reporting levels for the self-response treatment. However, their use of one-tailed significance tests is questionable, since two of the nonsignificant differences are in the direction of *more* reporting for the accept-proxy treatment.

As part of an investigation of increased use of the telephone for sample selection and interviewing in the NHIS, a recent experiment compared a randomly designated household respondent procedure with a self-selected "knowledgeable" household respondent procedure. Mathiowetz and Groves (1983) present treatment differences for the same set of indicators as Kovar and Wright (1973 op. cit.). Their findings, however, are quite the opposite of the earlier research. Most of the differences, including the only statistically significant effect, are in the direction of higher reported levels for the knowledgeable respondent (proxy) condition than for the self-responses from the randomly designated respondents. Within the random respondent treatment there was also a consistent trend for higher reporting levels among proxies. There is some evidence of the more "usual" trend for measures involving a longer recall period, but Mathiowetz and Groves summarize the results as showing an "overall tendency... directly counter to previous beliefs about self vs. proxy reports" (p. 96).

Several studies in the health area have controlled selection bias through a reinterview design, in which respondents originally interviewed by proxy are subsequently interviewed in person. The earliest of these occurred in conjunction with the California Health Survey. First, in a small pilot survey, 118 original proxy respondents were administered a self-response reinterview approximately two weeks after the initial interview. The reinterview was conducted without reference to the original interview, using the identical questionnaire. In general, the original proxy interview produced fewer reports of medical conditions than did the self-response reinterview. Discrepancies were substantial in both directions, however; each survey elicited many condition reports which had not been included in the other (California Department of Public Health (1957); Mooney (1962)).

A substantially larger reinterview program was also carried out in the subsequent main survey. For all of the health indicators examined – chronic conditions, acute conditions, restricted activity days, bed disability days – the self-response reinterview yielded a substantial increase in medical events and conditions over the original proxy reports. However, this increase is not totally attributable to self/proxy response status, since the reinterview also produced increased reporting among a sample of original self-respondents. The percentage increase from original interview to reinterview was greater for the original proxy group than for the original self-respondents, suggesting that response status contributed something to the difference (California Department of Public Health (1957); Mooney (1962)).

Koons (1973) has analyzed similar data from several years' reinterviews in the NHIS. His results also show increased reporting of health events and conditions in a self-response reinterview as compared to an original proxy interview, and the increases are typically larger than the interview-to-reinterview increases for those responding for self in the original interview. Thus, the results of these reinterview studies consistently suggest that the biases associated with self and proxy reporting of health conditions may differ, although differences between interview administrations may far outweigh any self/proxy differences.

In another health survey involving a type of reinterview design, Kolomel, Hirohula, and Nomura (1977) independently interviewed both members of 300 couples (mostly spouses) about each person's smoking, drinking, and diet behavior. The authors report great consistency between self and proxy reports, leading them to conclude that, for such inquiries, proxy reports are equal in quality to self-reports.

Only within the last few years have researchers begun to examine the effects of respondent status on income reporting. The 1979 panel of the Income Survey Development Program (ISDP) was the final major pretest in preparation for the Survey of Income and Program Participation (SIPP). This panel included an experiment on respondent rules, in which each sample household was randomly assigned to a maximum self-response treatment or an "accept-proxy" procedure.

Ferber and Frankel (1981) examine household ownership rates for 17 asset types (e.g., bonds, savings accounts, stocks, rental property, royalties, etc.) by respondent rule procedure. Differences in ownership rates are generally small, but consistently higher under the maximum self-response treatment. However, in a general summary of this test, Kulka (1982) concludes that the self/proxy response treatments produced no significant differences in asset reporting.

Income recipiency rates for both earned income and unearned government transfers also appear to have been unaffected by respondent rules. Kaluzny (1981a, 1981b) reports results from the same ISDP panel which show no significant differences for 13 income types. For respondents who reported any income, reported income amounts also did not differ.

Another self/proxy comparison in the ISDP is possible due to a special procedure in one wave of the 1979 panel – a self-response followup survey of students not living at home (and thus originally interviewed by proxy). Roman (1983) presents a comparison of self and proxy reports for 167 students who were successfully followed. The results suggest a higher rate of receipt of wage or salary income from the self-reports (66% versus 52% for proxy interviews), and interest income receipt (78% versus 66%), but no

significant differences in amount of educational assistance reported.

Martin and Butcher's (1982) self/proxy experiment, described earlier, also finds considerable consistency between self and proxy income reports. Of course, the proportion of reports "in agreement" depends on the definition of agreement: 81% of self-respondents and proxies placed gross weekly income in the same £50 category, 46% in the same £10 category, and less than a third in the same £1 category. Reported mean income differed only trivially, however – £74.8 per week for self-respondents versus £72.4 for proxies.

Hill (1987) offers a promising new technique for research on self/proxy issues. Using a modeling approach to control self-selection bias in uncontrolled self/proxy research, his results suggest that proxy reports of earnings income are substantially positively biased relative to self reports. Hill's work clearly underscores the dangers of drawing simple inferences from uncontrolled self/proxy studies. In this case the research bias is a misleading equivalence of mean self and proxy income reports (prior to the application of statistical controls) when – due to such sample differences as hours worked per week – there should be a difference.

Evidence on self/proxy report differences can also be found in Kinsey et al.'s (1948) investigation of sexual behavior, in which they conducted separate interviews with each spouse in a group of husband/wife pairs. The authors compare spouses' reports on 32 items, most of which involve relationship issues for which the self/proxy distinction is not appropriate. For seven of the items, however, one member of the pair is clearly the object of inquiry: husband's education, age at marriage, and occupation; and wife's education, age at marriage, number of abortions, and percentage of coitus with orgasm. None of these items shows a significant difference between the mean responses of husbands and wives.

Thus far, an examination of self/proxy report level differences from controlled treatment studies yields evidence of consistent effects only for reports of criminal victimization – and the latter is based on only one study. Labor force participation and income surveys have almost universally shown no significant treatment effects. (Hill's (1987) technique, which does indicate self/proxy bias differences for income reports, may now open up past and future uncontrolled treatment studies for more informative analysis.) This is also the most common result in health survey experiments, although some significant effects – albeit contradictory from one study to the next – have been found in the health area. The most appropriate general conclusion to draw from this type of research is that the weight of the evidence does not indicate consistent differences in the reporting levels of self and proxy respondents.

3.1.2. Response quality differences

Self/proxy differences in report levels at best only suggest differences reporting quality. In order to inform the issue with confidence, controlled studies need to include a direct assessment of response quality. Appropriately designed record check studies provide the most incontrovertible evidence of self/proxy response quality differences, but such studies are rare, and thus research employing more indirect indicators of response bias will also be examined.

Turner's (1972) crime survey respondent rule research described previously did not include an independent validity check. However, compared to self-respondents, household (proxy) respondents showed a greater tendency to distribute their victimization reports unevenly over the twelve-month reference period. Both self and proxy respondents tended to report more incidents

in the most recent six months of the reference period; for most incidents, however, this effect was more pronounced for the household respondent procedure. For all incidents combined, this treatment produced 50% more reports in the first half of the reference period than in the second half; for the self-response treatment the comparable figure was 41%. These differences suggest that for at least one component of response quality – the accurate dating of crime events, or their more complete recall – proxy responses may be more biased.

Deighton's (1967) comparison of a maximum self-response and a standard household respondent procedure indicated no reduction in the typical CPS "month-in-sample" bias (see Bailar (1975)) with a self-response procedure. In fact, the rotation group differences in the "in labor force" and the "unemployed" categories are more pronounced in the self-response treatment, although there are no statistically significant effects. Similarly, Aquilino (1971) presents results from an experimental panel of the CPS in which the data for all adults were obtained by self-response in the first month, and by both self and proxy response in the second month. The results show equivalent change in labor force classification from the first to the second month for self-self and self-proxy reports. Thus, although their quality assessment procedures are indirect, neither of these two investigations yields evidence of response bias differences between self and proxy labor force reports.

This review uncovered only two studies – both health surveys – in which controlled self/proxy response status was combined with an independent, comprehensive assessment of data quality. First, Cobb, Thompson, Rosenbaum, Warren, and Merchant (1956) summarize results of a three-phase investigation involving: (1) a household health interview, the first question of which asked whether any household member had arthritis or rheumatism, and if so, who; (2) an individual (self-response) interview with a subsample of persons from the household sample, using a questionnaire devoted solely to arthritis and rheumatism; and (3) a medical examination of a subsample of the self-respondents to detect either current symptoms or a history of arthritis or rheumatism.

Of the 707 persons selected for all three phases of the study, only 429 (61%) provided complete data, with most of the attrition due to refusal of the medical examination. However, the authors assert that "only minor differences were found in examination rate by age, sex, income group, history of arthritis, joint pain, or joint swelling and many other variables." and that the differences "are insufficient to affect the conclusions" (p. 135).

The medical examination placed each person in one of four groups: (1) definite arthritis or rheumatism; (2) symptoms of arthritis or rheumatism but no definite diagnosis; (3) not classifiable in (1) or (2) but without certainty that the person had never had arthritis or rheumatism; and (4) no suspicion of arthritis or rheumatism. (The original paper presents a five-category diagnostic scheme; for simplicity, two of the original categories – "classical arthritis" and "definite arthritis" – are combined here in (1).) Table 1 compares the arthritis or rheumatism prevalence estimates from the household and individual interview reports with the physician's diagnosis. Regardless of how one simplifies the medical diagnosis into a present/absent scheme – that is, whether (1) alone indicates the presence of arthritis or rheumatism, or (1) or (2), or any category but (4) – the bias difference between the reports obtained under the two procedures is very small.

Thompson and Tauber (1957) report results of a similar three-phase investigation of heart disease, involving a household inter-

Table 1. Estimates of arthritis or rheumatism based on a household (proxy) interview, an individual (self) interview, and a medical examination. Percent (Cobb et al. (1956))

Survey Prevalence rate		Examination Prevalence rate		Net bias (Survey – Examination)	
Household	Individual			Household	Individual
34.0	37.5	(1)*	34.7	– 0.7	+ 2.8
34.0	37.5	(1,2)*	51.0	–17.0	– 13.5
34.0	37.5	(1,2,3)*	76.7	–42.7	– 39.2

* Positive diagnosis criteria (see text).

Table 2. Estimates of heart disease based on a household (proxy) interview, an individual (self) interview, and a medical examination. Percent (Thompson and Tauber (1957))

Survey Prevalence rate		Examination Prevalence rate	Net bias (Survey – Examination)	
Household	Individual		Household	Individual
24.7	33.0	33.0	–8.3	0

Table 3. Estimates of heart disease based on a household interview, an individual interview, and a medical examination for persons interviewed for self in the household interview. Percent (Thompson and Tauber (1957))

Survey Prevalence rate		Examination Prevalence rate	Net bias (Survey – Examination)	
Household	Individual		Household	Individual
24.4	34.2	35.8	–11.4	–1.6

view, an individual interview, and a medical examination. Only about half of those sampled for the individual interview and examination actually completed all three phases of the study. Table 2 summarizes Thompson and Tauber's results for persons who were responded for by proxy in the household survey and who subsequently responded for themselves in the individual interview. This table appears to confirm the traditional assumption of proxy underreporting; additional evidence in Table 3, however, suggests another explanation.

Table 3 presents the household and individual interview results for persons who responded for themselves in the initial household interview. The household interview bias attributed to proxies in Table 2 is equally apparent in the data summarized in Table 3, where no proxies are involved. The "proxy bias" interpretation of Table 2 is not justified, since the same bias difference occurs

among those who self-responded in both interviews.

Perhaps the household interview suffered because it used only a single, global question to elicit reports of heart disease, whereas the individual survey was "oriented solely toward diseases of the heart" (Thompson and Tauber (1957, p. 1131)). Or perhaps the 18-month delay between the household interview and the medical examination (versus the week or so delay for the individual interview) was at fault; longer delay would increase the likelihood of real change in health status, and real change would almost certainly masquerade as underreporting. Other possibilities may also exist, but whatever the cause of the Thompson and Tauber findings, it is not likely to have been "proxy bias."

Thus, inspection of the relevant research does not reveal strong or consistent evidence to support the notion that proxy data are in general more biased than self-response data. The results of a single study suggest a slight tendency for self-respondents to distribute crime incident reports more evenly across a 12-month reference period than do proxies. However, two investigations of labor force reporting show no bias differences due to response status, and two health studies, which included direct checks of response validity, also fail to support the assumption that self-reports are less biased than proxy reports.

3.2. *Response error variance effects*

Another important dimension of data quality is the extent to which respondents reply accurately, regardless of the direction of their errors. A procedure may yield unbiased estimates without producing a single accurate reply (as long as the errors are perfectly compensating), and a more biased procedure may produce a greater number of accurate replies than a less biased one. If first-order point estimates are all that is of interest, then bias is the only component of data quality that need be considered. If, however, the higher-order aspects of the data are of interest (e.g., transition estimates, multivariate relationships, etc.), then response error variance is also important.

As was the case for response bias, sound research on response error variance is rare and often employs only indirect measures of quality. In fact, the clearest data on response error variance effects are from the two health studies described in the previous section.

Table 4 summarizes the Cobb et al. (1956) arthritis and rheumatism results in terms of the gross accuracy rate – the proportion of survey responses which agree with the medical examination – for the household (proxy) and individual (self) interviews. Regardless of how the medical diagnosis of arthritis or rheumatism is defined, the gross accuracy rate difference between the household and the individual interview is trivial.

Table 4. Proportion of household (proxy) and individual (self) interview responses regarding arthritis or rheumatism agreeing with a medical examination. Percent (Cobb et al. (1956))

Medical examination categories indicating a positive diagnosis (see text for explanation)	Gross accuracy rate	
	Household	Individual
(1)	71.8	71.6
(1,2)	73.2	73.2
(1,2,3)	55.9	59.9

Table 5. Proportion of household and individual interview responses regarding heart disease agreeing with a medical examination, for all respondents and for those responded for by a proxy in the household interview. Percent (Thompson and Tauber (1957))

Respondent group	Gross accuracy rate	
	Household	Individual
All respondents	77.3	76.8
Proxy in household interview	73.1	72.5

Thompson and Tauber's (1957) data, summarized in Table 5, point to the same conclusion. The proportion of accurate replies is virtually identical in the household and individual interviews, even restricting consideration to just the subset of respondents for whom the initial household interview was actually a proxy interview.

The previously described experimental panel of the Income Survey Development Program (ISDP) included a respondent rules experiment, comparing a maximum self-response procedure with an accept-proxy procedure. Evaluations of this experiment have employed various indirect indicators of quality (such as the extent of rounding of income amounts, variances of income amounts, and the respondent's use of records to assist accurate reporting of income amounts) related to the random error dimension of quality.

Kaluzny (1981a, 1981b) presents results indicating some tendency for more reporting of rounded (i.e., divisible by 5) income amounts under conditions more tolerant of proxy response, but the differences are not consistent across all income types nor within the same type across survey waves. Differences in the variances of reported income amounts are also inconsistent. Income amount variances are generally lower for the self-response treatment in wave 1 of the 1979 ISDP panel, but the wave 2 results show an equal number of differences in both directions. The only clear evidence in the ISDP of a quality difference favoring the self-response procedure is in the respondent's use of records (Kaluzny (1981a); Vaughan (1980)).

As noted in the preceding section, a recent paper by Hill (1987) uses a modeling approach to control self-selection bias in uncontrolled self/proxy research. Hill's results suggest substantially greater response variance for proxy reports of earned income than for self-reports. However, this difference appears due to a few extreme cases; removing these outliers reverses the original difference, resulting in significantly lower variance for proxy reports. Martin and Butcher (1982) report a slightly higher variance for proxy income reports in their self/proxy, dual interview study, although for respondents of higher "social class" or educational attainment this trend is reversed.

Once again, the limited research evidence fails to support the hypothesis that proxy data are inferior to data obtained by self-response – specifically, that they are beset with greater numbers of inaccurate replies. Two health studies show virtually identical levels of response accuracy for self and proxy reports. Attempts to find quality differences (with several different quality indicators) in income reporting show only weak and inconsistent effects, with the exception of the respondent's use of records to assist accurate recall.

3.3. *Nonresponse effects*

Response status may also affect data quality through its effects on response completeness – item nonresponse, person nonresponse, or household nonresponse. Results from two labor force surveys suggest that self-response rules may produce less complete data. Deighton (1967) reports a slightly (but not significantly) higher household noninterview rate with a self-response procedure – 6.2%, versus 5.9% for the standard household respondent treatment – and some person noninterviews in interviewed households (0.6%) where the standard treatment had none. Singh and Tessier (1975) report even more dramatic results in an experimental self-response panel of Canada's Labour Force Survey. The self response panel had a household noninterview rate of about 11%, versus only 6% in the household respondent parent survey.

Kovar and Wright (1973) found no differences in household nonresponse between an experimental self-response treatment and the standard (accept-proxy) NHIS response rules. However, they do report a very small increase in person nonresponse (of about 1%) under self-response conditions. Kovar and Wilson (1976) suggest that this latter effect may only apply to males; regardless, it is probably too small to be of practical significance.

The experiment conducted in the 1979 panel of the ISDP has generated extensive investigations of self/proxy nonresponse effects. For example, Vaughan (undated (a); undated (b)) presents data from the first two (of six) waves of the 1979 panel suggesting slightly higher household refusal and total noninterview rates for the self-response treatment. Olson (1981) corroborates this tendency for all waves of the 1979 panel, but concludes that "attrition differences by treatment are too small to give guidance in the self/proxy decision" for the SIPP itself (p. 1). The most interesting aspect of these results is that the differences appeared as early as the first survey wave, before respondents had any opportunity to be affected by the self-response procedures. This suggest that noninterview rate differences favoring the accept-proxy procedure might better be attributed to interviewer reluctance to administer the self-response procedure than to any negative reaction from respondents.

Similar slight differences are evident in person noninterview rates within interviewed households. Vaughan (1980) reports a 2.5% noninterview rate with the maximum self-response treatment in the first ISDP wave, versus 1.2% under the standard accept-proxy treatment. Vaughan characterizes this difference as statistically but not practically significant (p.2).

Much attention has been directed toward comparisons of item nonresponse for the two respondent rule treatments in the ISDP. This attention undoubtedly reflects the great concern with nonresponse rates for income items in the traditional major sources of income data in the United States – the decennial census and the March CPS income supplement. Coder (1980) presents first wave nonresponse rates for six income items in the 1979 panel: hourly wage rate, Social Security Income, Federal Supplemental Security Income, pension and retirement income, self-employment income, and rental income. In each case, item nonresponse under proxy conditions is substantially higher than under the self-response rule. Vaughan (1980) reports nonresponse information for six income variables, which also show consistently greater nonresponse under the accept-proxy treatment.

Both Coder and Vaughan used preliminary and unedited data, which may explain why they are at odds with Kaluzny's (1981a, 1981b) later examination of item nonresponse on amounts received from seven

income sources. Kaluzny's results present a much less consistent picture. In wave 1 only four of the seven income types show greater item nonresponse under accept-proxy conditions, and most of the differences are reversed in wave 2. Kulka (1982) summarizes the various investigations as not having demonstrated consistently lower item nonresponse rates under self-response, with the possible exception of hourly wage and Social Security Number reporting. Martin and Butcher's (1982) data, however, support the more usual trend – for all of the labor force and income items examined, proxy item nonresponse (including "don't know" and "uncodeable" replies) exceeded that of the self-respondents.

Although the effects are neither large nor consistent, the research evidence suggests that self/proxy status may have some reliable effects on response completeness. Across subject-matter areas, it appears that self-response procedures produce higher household and person noninterview rates. Overall response completeness may be equivalent, however, since these differences seem to be balanced by lower item nonresponse under procedures which maximize self-response.

4. Summary and Conclusions

This review of the literature finds little support for the notion that self-response survey reports are of generally better quality than proxy reports. In practical terms, this suggests that survey designers should use self-response interviews if they are easily obtainable, but need not undertake extraordinary efforts to maximize self-response. Crime studies may be an exception to this generalization. The existing data should give researchers some confidence that the responses of proxies are comparable in quality to what would have been obtained via self-response with additional effort.

The more obvious conclusion is that there is really not enough evidence to draw solid conclusions. Well designed studies of the self/proxy issue are very rare, and the range of subject matter covered has been limited. Furthermore, only the most basic overall estimates have received any attention. It is quite possible, for example, that self/proxy status may have negligible general effects, and yet may interact with respondent characteristics so as to significantly affect age-specific (or other) estimates. Martin and Biderman (1984) and O'Muircheartaigh (1986) present data which suggest this possibility.

So, this review, too, must conclude with a call for more – and better – research. Too much of the work that has been done has been relatively easy and inexpensive, but has not really advanced our knowledge or provided practical guidance for survey planners. A sound research program would provide the information for more rational decision-making.

5. References

Aquilino, R. (1971): Methods Test Phase III: Third Report on the Accuracy of Retrospective Interviewing and Effects of Change in Respondent on Labor Force Data. U.S. Bureau of the Census memorandum (April 2, 1971), Washington, D.C.

Bailar, B.A. (1975): The Effects of Rotation Group Bias on Estimates from Panel Surveys. Journal of the American Statistical Association, 70, pp. 23–30.

Berk, M.L., Horgan, C.M., and Meyers, S.M. (1982): The Reporting of Stigmatizing Health Conditions: A Comparison of Proxy and Self-reporting. American Statistical Association, Proceedings of the Section on Social Statistics, pp. 506–509.

California Department of Public Health (1957): Health in California. Department of

Public Health, Berkeley, CA.

Cartwright, A. (1957): The Effect of Obtaining Information from Different Informants on a Family Morbidity Inquiry. Applied Statistics, 6, pp. 18–25.

Cartwright, A. (1963): Memory Errors in a Morbidity Survey. Millbank Memorial Fund Quarterly, 41, pp. 5–24.

Cobb, S., Thompson, D.J., Rosenbaum, J., Warren, J.E., and Merchant, W.R. (1956): On the Measurement of Prevalence of Arthritis and Rheumatism from Interview Data. Journal of Chronic Diseases, 3, pp. 134–139.

Coder, J.F. (1980): Some Results from the 1979 ISDP Research Panel. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 540–545.

Commission on Chronic Illness (1957): Chronic Illness in a Large City: The Baltimore Study. Harvard University Press, Cambridge, MA.

Cowan, C.D., Roman, A.M., Wolter, K.M., and Woltman, H.F. (1979): A Test of Data Collection Methodologies: The Methods Test. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 141–146.

Deighton, R.E. (1967): Some Results of Experimentation with Self-Respondent Interviewing Procedures, February 1965 – June 1966. U.S. Bureau of the Census memorandum (February 28, 1967), Washington, D.C.

Elinson, J. and Trussell, R.E. (1957): Some Factors Relating to Degree of Correspondence for Diagnostic Information as Obtained by Household Interviews and Clinical Examinations. American Journal of Public Health, 47, pp. 311–321.

Enterline, P.E. and Capt, K.G. (1959): A Validation of Information Provided by Household Respondents in Health Surveys. American Journal of Public Health, 49, pp. 205–212.

Ferber, R. (1955a): On the Reliability of Purchase Influence Studies. Journal of Marketing, 19, pp. 225–232.

Ferber, R. (1955b): On the Reliability of Responses Secured in Sample Surveys. Journal of the American Statistical Association, 50, pp. 788–810.

Ferber, R. and Frankel, M. (1981): Evaluation of the Reliability of the Net Worth Data in the 1979 Panel: Asset Ownership on Wave 1. University of Illinois, Survey Research Laboratory, Champaign-Urbana, IL.

Haase, K.E. and Wilson, R.W. (1972): The Study Design of an Experiment to Measure the Effects of Using Proxy Respondents in the National Health Interview Survey. American Statistical Association, Proceedings of the Social Statistics Section, pp. 289–293.

Haberman, P.W. and Elinson, J. (1967): Family Income Reported in Surveys: Husbands Versus Wives. Journal of Marketing Research, 4, pp. 191–194.

Hansen, M.H., Hurwitz, W.N., Marks, E.S., and Mauldin, W.P. (1951): Response Errors in Surveys. Journal of the American Statistical Association, 46, pp. 147–190.

Hill, D.H. (1987): Assessing the Relative Quality of Proxy Data from a Non-Experimental Study. Paper presented at the Third Annual Research Conference (March, 1987), U.S. Bureau of the Census, Washington, D.C.

Horvitz, D.G. (1952): Sampling and Field Procedures of the Pittsburgh Morbidity Survey. Public Health Reports, 67, pp. 1003–1012.

Jones, C. and Aquilino, R. (1970): Methods Test Phase III: Second Report on the Accuracy of Retrospective Interviewing and Effects of Nonself Response Labor Force Status. U.S. Bureau of the Census memorandum (January 29, 1970), Washington, D.C.

Kaluzny, R.L. (1981a): Evaluation of Experi-

mental Effects, 1979 Research Panel – Wave 1. Mathematica Policy Research, Inc., Princeton, NJ.

Kaluzny, R.L. (1981b): Evaluation of Experimantal Effects, 1979 Research Panel – Wave 2. Mathematica Policy Research, Inc., Princeton, NJ.

Kilss, B. and Alvey, W. (1976): Further Exploration of CPS-IRS-SSA Wage Reporting Differences for 1972. American Statistical Association, Proceedings of the Social Statistics Section, pp. 471–476.

Kinsey, A.C., Pomeroy, W.B., and Martin, C.E. (1948): Sexual Behavior in the Human Male. Saunders, Inc., Philadelphia, PA.

Kohn, M.L. and Carroll, E.E. (1960): Social Class and the Allocation of Parental Responsibilities. Sociometry, 23, pp. 372–392.

Kolomel, L.N., Hirohula, T., and Nomura, A. (1977): Adequacy of Survey Data Collected from Substitute Respondents. American Journal of Epidemiology, 106, pp. 476–484.

Koons, D.A. (1973): Quality Control and Measurement of Nonsampling Error in the Health Interview Survey. National Center for Health Statistics Pub. No. (HSM) 73–1328, Rockville, MD.

Kovar, M.G. and Wilson, R.W. (1976): Perceived Health Status – How Good is Proxy Reporting? American Statistical Association, Proceedings of the Social Statistics Section, pp. 495–500.

Kovar, M.G. and Wright, R.W. (1973): An Experiment with Alternate Respondent Rules in the National Health Interview Survey. American Statistical Association, Proceedings of the Social Statistics Section, pp. 311–316.

Krueger, D.E. (1957): Measurement of Prevalence of Chronic Disease by Household Interviews and Clinical Evaluation. American Journal of Public Health, 47, pp. 953–960.

Kulka, R.A. (1982): Tests and Experiments. Chapter 4 in Research Triangle Institute, ISDP 1979 Research Panel Documentation, Research Triangle Institute, Research Triangle Park, NC.

Levinger, G. (1966): Systematic Distortion of Spouses' Reports of Preferred and Actual Sexual Behavior. Sociometry, 29, pp. 291–299.

Marquis, K.H. (1983): Record Checks for Sample Surveys. Paper presented at the Advanced Research Seminar on Cognitive Aspects of Survey Methodology (June, 1983), Committee on National Statistics, National Research Council, Washington, D.C.

Marquis, K.H., Duan, N., Marquis, M.S., and Polich, J.M. (1981): Response Errors in Sensitive Topic Surveys. Rand Corporation, Santa Monica, CA.

Martin, E.A. and Biderman, A.D. (1984): Memorandum on Proxy vs. Self-response for 12–13 Year-olds. Crime Survey Research Consortium Items 101–141, Bureau of Social Science Research, Washington, D.C.

Martin, J. and Butcher, B. (1982): The Quality of Proxy Information – Some Results from a Large-scale Study. The Statistician, 31, pp. 293–319.

Mathiowetz, N. and Groves, R.M. (1983): The Effects of Respondent Rules on Health Survey Reports. Chapter V in C.F. Cannell et al., An Experimental Comparison of Telephone and Personal Health Surveys, National Center for Health Statistics, Rockville, MD.

Mooney, H.W. (1962): Methodology in Two California Health Surveys: San Jose (1952) and Statewide (1954–55). Public Health Monograph No. 70, Government Printing Office, Washington, D.C.

Mudd, E.H., Stein, M., and Mitchell, H.E. (1961): Paired Reports of Sexual Behavior of Husbands and Wives in Conflicted Marriages. Comprehensive Psychiatry, 2, pp. 149–156.

Neter, J. and Waksberg, J. (1965): Response Errors in Collection of Expenditures Data

by Household Interviews: An Experimental Study. Technical Paper No. 11, U.S. Bureau of the Census, Washington, D.C.

Nisselson, H. and Woolsey, T.D. (1959): Some Problems of the Household Interview Design for the National Health Survey. Journal of the American Statistical Association, 54, pp. 69–87.

Olson, J. (1981): Self-Proxy Treatment Rules and Panel Attrition. U.S. Department of Health and Human Services memorandum (April 20, 1981), Washington, D.C.

O'Muircheartaigh, C. (1986): Correlates of Reinterview Response Inconsistency in the Current Population Survey (CPS). Paper presented at the Second Annual Research Conference (March 1986), U.S. Bureau of the Census, Washington, D.C.

Radke, M.J. (1946): The Relation of Parental Authority to Children's Behavior and Attitudes. University of Minnesota Press, Minneapolis, MN.

Roman, A.M. (1983): Preliminary Findings from Comparisons of Self Versus Proxy Response for Students from the 1979 ISDP. U.S. Bureau of the Census memorandum (September 29, 1983), Washington, D.C.

Roman, A.M. and Woltman, H.F. (1980): The Methods Test Panel: Analysis of the Unemployment Rate: An Interim Report: June 1978 – March 1979. U.S. Bureau of the Census report (March 14, 1980), Washington, D.C.

Roshwalb, A. (1982): Respondent Selection Procedures Within Households. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 93–98.

Rutter, M. and Brown, G.W. (1966): The Reliability and Validity of Measures of Family Life and Relationships in Families Containing a Psychiatric Patient. Social Psychiatry, 1, pp. 38–53.

Singh, M.P. and Tessier, R. (1975): MTP (Methods Test Panel) Phase III: Feasibility Test for Complete Non-proxy Procedure, Methodology and Analysis. Statistics Canada, Ottawa, Canada.

Sudman, S. and Bradburn, N.M. (1974): Response Effects in Surveys. Aldine, Chicago, IL.

Thompson, D.J. and Tauber, J. (1957): Household Survey, Individual Interview, and Clinical Examination to Determine Prevalence of Heart Disease. American Journal of Public Health, 47, pp. 1131–1140.

Turner, A.G. (1972): Methodological Issues in the Development of the National Crime Survey Panel: Partial Findings. Law Enforcement Assistance Association unpublished report (December 1972), Washington, D.C.

Turner, C.F. and Martin, E.A. (1984): Surveying Subjective Phenomena. Basic Books, New York, NY.

Vaughan, D. (1980): Preliminary Findings on the Data Quality Effects of Employing Self-response Rules in the 1979 ISDP Panel. Social Security Administration draft memorandum (February 16, 1980), Washington, D.C.

Vaughan, D. (undated a): Interview Visits by Response Mode and Form Type, February Interview, 1979 ISDP Panel. Social Security Administration draft memorandum, Washington, D.C.

Vaughan, D. (undated b): Some Evidence on Response Mode and Refusal Rates, Wave One and Two of the 1979 ISDP Panel. Social Security Administration draft memorandum, Washington, D.C.

Williams, L.E. (1969): Methods Test Phase III: First Report on the Accuracy of Retrospective Interviewing and Effects of Non-self Response on Labor Force Status. U.S. Bureau of the Census memorandum (June 24, 1969), Washington, D.C.

Received April 1987
Revised December 1987

Journal of Official Statistics
Vol. 4, No 2, 1988, pp. 173 – 190

Trends in Agricultural Statistics – An Outline of Development Work at Statistics Sweden

Knut Medin[1] *and Bernt Wilson*[2]

Abstract: In recent years, Statistics Sweden has conducted a general review of its agricultural statistics. This article emphasizes the prerequisites and motives for this development work. Stress is laid on a statistical approach, and on improvements in statistical methodology.

Among the methodological projects mentioned are the use of a farm typology, the idea of farm models, longitudinal and spatial analyses, projections and forecasts, and the use of geographical coordinates.

Examples of "new" statistical areas include agricultural structure and infrastructure, family farming, and part-time farming. Respondent burden, confidentiality and integrity, and the demands of budgetary restraints are discussed as important background factors.

Key-words: Agriculture; official statistics; Sweden; development work; statistical methodology.

1. Introduction

In Sweden those parts of official agricultural statistics that refer to farms are, in most cases, produced by Statistics Sweden. This article describes current developments within this branch of statistics, but the objective also is to examine the background of the development work and especially to indicate prerequisites and motives.

The article includes: a short overview of Swedish agricultural statistics; a discussion on needs, aims, and restrictions of the development work; and a survey of development projects.

Some of these projects, such as the one referring to "spatial analysis," are of a general methodological character. Most of the projects, however, deal with subject matter. In this case the article concentrates on areas which can be regarded as more or less "new," at least for Swedish agricultural statistics.

From a historical point of view, it can be said that modern agricultural statistics was introduced when the Central Bureau of Statistics (now called Statistics Sweden) started planning its first Census of Agriculture, taken in 1927. Since then development work has continued with varying intensity, directed both by progress in statistical methodology and by

[1] Consultant to Statistics Sweden, Stockholm; formerly Head of the Department for Area Statistics.
[2] Head of the Division for Postal Surveys in Agriculture, Statistics Sweden, Örebro.

Acknowledgments: The authors wish to acknowledge the assistance given by numerous colleagues at Statistics Sweden. This also includes many helpful comments and suggestions made in connection with earlier versions of the article. The authors also wish to thank the referees for their valuable suggestions.

the changing need for agricultural statistics for political, administrative, and scientific purposes.

The present phase of Swedish agricultural statistics goes back to the late 1960s, when a government committee proposed a number of changes. The most important change was the introduction of the Farm Register, which now has a central role in the overall system of agricultural statistics.

The development work dealt with here started in the middle of the 1970s and was directed by a small working group. The group delivered two main reports, see Statistics Sweden (1978, 1985a), which have been summarized by Hedqvist and Thorburn (1978) and by Hedqvist and Rösiö (1985). About 80 changes and development projects were proposed by the working group. Most of these have now been completed.

The authors of this article have been members of the working group; this article is freely based on the group's reports.

2. Swedish Agricultural Statistics – An Overview

Major activities in agricultural statistics at Statistics Sweden are presented in Table 1.

Official agricultural statistics that are based on nonfarm data, i.e., data from mills, slaughterhouses, producers of agricultural machinery, etc., are in most cases produced by the National Agricultural Market Board. That board is also responsible for various forecasts and sector analyses.

In Sweden, forestry statistics are not regarded as a part of agricultural statistics but as a statistical branch of its own, for which the National Board of Forestry bears the main responsibility. However, because of the close links in this country between agriculture and forestry, these two economic activities can not always be separated in statistical reporting.

Outside the scope of agricultural statistics (though in most cases produced by Statistics Sweden) are official statistics in which agriculture represents only one of many activities, such as population statistics, labour force statistics, national accounts, etc.

In addition to the brief descriptions of the various activities presented in Table 1, some further comments may be appropriate with regard to the Farm Register, the Objective Crop Yield Surveys, and the Farm Bookkeeping Survey.

A major part of Swedish agricultural statistics is based on the Farm Register. Since it was established in 1968, the register, in addition to statistical purposes, has been used administratively, inter alia by the 24 County Agricultural Boards and in the administration of the crop insurance.

The register consists of all enterprises (i.e., holdings, farms) in agriculture, horticulture, and forestry. It contains about 300 000 units which are divided into three categories:

1. about 105 000 holdings with at least 2.0 hectares of arable land;
2. about 2 000 additional holdings specializing in horticulture or with large-scale animal husbandry; and
3. about 190 000 other holdings, most of which with forest land only.

Data on categories (1) and (2) are brought up to date by an annual postal inquiry. Data on category (3) are brought up to date at an interval of about five years by the use of official real estate taxation registers.

The annual data collection takes place in the middle of June. It is compulsory for the farmer to complete a four-page form, which is distributed to him or her on the basis of the previous year's register. The Farm Register is in this way updated with regard to information on the holders, the real-estate units, acreages by different types of land, use of

Table 1. Major activities in agricultural statistics at Statistics Sweden

Farm Register (annual updating)	Registered data used for current statistics on holdings, holders, crop and animal husbandry, horticulture, etc. Also for censuses of agriculture and longitudinal studies. Sampling frame for regular and ad hoc sample surveys. Register created especially but not exclusively for statistical purposes (used for the calculation of crop damage compensations, etc.). Details for each farm updated by postal survey. All farms.
Census of Horticulture (every third year)	Statistics on plants grown, areas under glass or in the open, quantities produced, heating technique, etc. Content varies somewhat as between censuses. Postal survey (in some cases interviews). All 5 000 horticultural holdings.
Objective Crop Yield Surveys (annual, continous during season)	Yield per hectare statistics for eight important field crops (with detailed regional breakdown). In addition certain other statistics on crop husbandry. Survey based on the collection of crop samples combined with interviews. 12 000 farms.
Crop Outlook Reports (three times a year during season)	Crop outlook assessments made for 2 400 districts (whole of Sweden) by local agents. Basis for quantitative forecasts.
Farm Bookkeeping Survey (annual)	Statistics on farm economics in monetary and physical terms, calculated results. Data collection (whenever possible based on farmer's own bookkeeping) through local accounting offices. 1 000 farms.
Survey of Farmers' Assessed Incomes, Expenditures, Net Earnings, etc. (annual)	Statistics based on data reported by farmers to taxation authorities. Data transcribed and edited at Statistics Sweden. 7 000 farms. In addition a limited set of statistics based on data for all farms are obtained from the EDP-based tax assessment register combined with the Farm Register.
Agricultural Labour Force Survey (annual, as from 1988 every third year)	Telephone interviews four times a year. 1 000 farms.
Statistics on Boars and Covered Sows (monthly)	Data collection from all 6 000 boar keepers by postal survey.
Survey of Building Activities on Farms (annual)	Statistics on the construction and reconstruction of farm buildings. Telephone interviews based on a preliminary data collection linked to the farm registration.

arable land for different crops, horticultural production, and numbers of certain animals. In addition, there usually are some special items that vary from year to year.

Almost all data collected for the farm register is used for statistical purposes. The register is also used as a sampling frame for most regular and occasional sample surveys in agriculture. Since the concepts, definitions, identity numbers, etc., are the same in the register and in the sample surveys, it has been feasible to build up a highly consistent system for agricultural statistics. The Farm Register makes it possible: (1) to follow each holding over time, (2) to check the coverage against other official registers, especially those regarding land area, (3) to produce statistics for small administrative regions (parishes, municipalities, etc.), and (4) to produce new agricultural statistics without increasing the volume of data collected, for instance by new cross-classifications or through longitudinal studies.

Technically the register has been continually improved in recent years, especially by replacing manual methods by more efficient electronic data processing (EDP) methods. For details about the Farm Register, see Medin (1985) and the references presented there; see also Statistics Sweden (1983).

Since about 1960, the official statistics on crop yields are for main crops – winter wheat, spring wheat, rye, barley, oats, grass, table potatoes, and potatoes for processing – obtained through the Objective Crop Yield Surveys covering all of Sweden.

A sample of about 12 000 farms is selected annually from the Farm Register. On each of these farms, one field is selected for each crop included in the surveys, provided that crop is cultivated on that particular farm. On that field one to three sample plots are selected.

For grains and grass each plot is a circle one square metre in area; for potatoes the plot is two metres in length along a row. These plots are harvested shortly before the farmers' own harvest by local personnel employed by the County Agricultural Boards. For grains and grass, the plot crop[3] is sent to a central laboratory for drying and weighing. The potato crop is weighed in the field.

In this way, data on the biological yield are obtained. In addition subsamples are used to measure harvesting losses, yield quality, use of different crop varieties, farming practices, etc. The estimates are said to be objective since crop cutting is applied in combination with random sampling procedures.

The large sample size is needed to provide estimates for the 420 crop yield districts used for crop insurance calculations. Of course the results are used for many other purposes as well.

Doubts about the justification of this type of survey have been expressed by, e.g., Zarkovich (1977). However, in Sweden, the reliability of the survey results has not been called into question in recent years (for sufficiently large samples).

The main structure of the surveys has not changed much over the last 10 – 15 years but many details have been modified. Cost savings in both the sampling procedures and in the field work have been introduced. For details about the Objective Crop Yield Surveys, see Statistics Sweden (1987c) and Medin (1965). Söderlind (1982) documented the development of the surveys. For references to early papers by Nilsson, Zetterberg, and Söderlind, see Zarkovich (1977). The sampling design was studied by Jönrup (1976), among others.

For more than 70 years farm bookkeeping data has been systematically used to study

[3] The crop quantity taken away is so small that it has not been regarded as necessary to compensate the farmer.

the economics of Swedish farming. At the beginning, this data collection could be characterized as scientific research in agricultural economics. Because of the increased use of the results as a basis for the agricultural policy debate the Farm Bookkeeping Survey has been gradually transformed into official statistics.

In 1976 the Farm Bookkeeping Survey was transferred from the National Board of Agriculture to Statistics Sweden. Since then, the survey has been closely coordinated with agricultural statistics in general. A sample design based on the new farm typology is now being introduced. Because of the high costs per unit, the Farm Bookkeeping Survey comprises only about 1000 farms chosen from a population consisting of certain farm categories of particular interest for agricultural policy deliberations. Every farm selected participates in the survey for four years, and one quarter of the sample is renewed every year.

The results tabulated in the Farm Bookkeeping Survey cover a large number of variables, both physical and monetary. In the last few years various models for calculating profitability given inflation have also been developed. For details about the survey, see Larsson, Medin, and Wilson (1987).

In the collection of farm data, Statistics Sweden has in recent years been assisted by regional agricultural authorities and organizations. The most important of these cooperating bodies have been the 24 County Agricultural Boards, which fall under the National Board of Agriculture. They take part in data collection not only for the Objective Crop Yield Surveys, but also for the Farm Register and for several other subbranches of agricultural statistics. In the data collection for the Farm Bookkeeping Survey, Statistics Sweden cooperates with local offices of a nation-wide accounting organization, belonging to the Federation of Swedish Farmers.

3. Need for Updating Procedures and Statistics

3.1. Developments in the farming industry

Needless to say, agricultural statistics have played a considerable role as a source of information on the great changes that have taken place in Swedish farming, especially after World War II. Intensive work was required to adjust statistical reporting to the changing technical, economic, and social conditions for farming and for rural life in general. Therefore, a few words on the development of Swedish agriculture may be appropriate.

Most agricultural products come from the plains in the southern and central parts of the country. However, there is a comparatively large number of small farms situated in the rest of Sweden, which is mainly forested. In recent years many small units have fallen into disuse or have been amalgamated into larger farms. Thus the total number of holdings decreased by about two-thirds from 1951 to 1986, while the total area of arable land decreased only from 3.5 to 2.9 million hectares. The most common type of farm enterprise is the family farm where almost all work is performed by the farmer and his or her family.

A decreasing need for manual work on farms, mainly due to mechanization, has led to an increase in the number of part-time farms. The farmer and spouse obtain more than half their total income from activities outside agriculture and forestry for about 65 % of all farms. Demand for efficiency has also led to more specialized production on farms.

3.2. Agricultural policy aspects

For a long while, the government has been deeply concerned about agriculture. Programmes approved by Parliament have had such aims as reasonable incomes for farmers,

reasonable food prices for consumers, and a level of food production in the country that guarantees a certain degree of self-suffiency in case of war. These aims should be achieved by rationalization measures, price regulations, import-export regulations, subsidies, etc.

For many different reasons the exact content of the policy has changed from time to time with major revisions at intervals. However, there has been an obvious ambition from the side of the government to base changes on relevant and accurate statistics. This extends to statistics for the administration of established programmes and also to follow-up activities.

A few current political issues with (possible) effects on the need for agricultural statistics are: surplus of farm products, consequences of present land use restrictions, food quality, and distribution of income between categories of farmers.

3.3. *Importance of the crop insurance*

The introduction of a national crop insurance scheme in 1961 was to have considerable effects on official agricultural statistics. This followed from the basic assumption that, for each crop, the loss per hectare is approximately equal for all farms within reasonably small districts. The loss is expressed as the difference between the "normal" yield value and the "actual" yield value.

For most of the important crops, district data on the average yield per hectare is obtained through the Objective Crop Yield Surveys; for other crops various methods are applied. Areas of different crops reported to the Farm Register are also used when calculating the indemnities to which farmers may be entitled.

Statistics Sweden has been responsible not only for the yield surveys and the Farm Register but also for the technical administration of the insurance. This administration has included the indemnity calculations for individual farmers although not the payments.

In connection with the crop insurance, Statistics Sweden has had to meet strict requirements for accurate basic data, regionalized statistics, timely EDP operations, etc. Necessary organizational and financial resources have, however, been made available with good results also for other uses of agricultural statistics. A description of the crop insurance during its early years was given by Medin (1965); an up-to-date review has been published by Statistics Sweden (1988).

According to a recent agreement between the government and the Federation of Swedish Farmers a new crop insurance scheme will be introduced in 1988. The main responsibility for the new insurance will be transferred to the federation. At the time this article was written, the full implications of these changes for official agricultural statistics were not yet apparent. It was, however, known that the Farm Register will be retained. The Objective Crop Yield Surveys will also be continued although in a considerably reduced form.

3.4. *International requests and foreign examples*

Sweden participated in the preparatory work for the first World Census of Agriculture around 1930 and later took active part in international efforts towards comparability and methodological development in agricultural statistics. Special reference should be made here to the work by the Food and Agricultural Organization, the Conference of European Statisticians (reporting to the Economic Commission for Europe) and the Organization for Economic Cooperation and Development as well as the work directed by the Chief Statisticians of the Nordic Countries. Agricultural statistics in the European

Economic Community have been given special attention and detailed comparisons can be found in Statistics Sweden (1976b).

4. Impact of Statistical Aims and Innovations

4.1. A statistical approach

In recent years, there has been a considerable discussion in Sweden about the significance of a "statistical approach" in official statistics. The discussion started with the presentation of a paper prepared in close connection with development work in agricultural and related statistics. It resulted in a special issue of *Statistical Review*, see Borglund, Jorner, Medin, Olofsson, and Polfeldt (1984).

4.2. The need to promote studies of change

Agricultural statistics traditionally reported estimates for a particular point in time or a particular time interval. The statistics have been based on data collected at more or less regular intervals. In addition to other applications, the results have sometimes been used for simple time series analyses.

Since changes in the agricultural industry, especially from the structural and economic points of view, have been greater and more rapid in recent decades, the need to analyze the changes, their causes and the change mechanisms has increased. This was stressed by Widén and Åstrand (1975).

Among other things, this interest in change has been reflected in demands for better comparability over time in statistical series. As a consequence, it has become necessary to clarify the effects of shifts and modifications in definitions, concepts, and methods used.

Longitudinal analyses (cf. Section 6.3) have been used to describe statistically the changes from time to time in different objects such as holdings, herds, etc. The "net change statistics" have in this way been supplemented with "gross change statistics."

4.3. Some quality aspects

This is not the place to deal with all aspects of quality in Swedish agricultural statistics but to highlight our experience.

It is sometimes argued that the quality of the statistics will be impaired if the statistical data collection is coordinated with a collection of data for administrative purposes. However, generally speaking this has not happened to the Farm Register. One reason for this may be that the farmers, when filling in the forms, are well aware of the fact that the data will be used for many different purposes. The farm register has also escaped the problems encountered when the collection of data for statistics has been added to a data collection routine already used by an administrative body.

An elaborate evaluation study[4] is the Annual Area Checking Survey in which the crop areas reported by farmers to the Farm Register are compared with control measurements in the fields, see Polfeldt (1977). This survey gives general information on the quality of the area data in the register. The results are also used for the calculation of correction factors to be applied when estimating the total yield of different crops. Furthermore, results for individual farms are used to check the data base for the crop insurance; in the case of a major error, legal action may be initiated by Statistics Sweden.

In almost all areas of agricultural statistics, comprehensive EDP editing programs have been applied. In some cases a technique called "macro checks" has been found a very

[4] As a historical note it may be mentioned that funds were explicitely made available for a separate evaluation study in connection with the first Swedish Census of Agriculture in 1927, as reported by Statistics Sweden (1936).

efficient complement to traditional checking. In macro checks, preliminary statistical summaries are edited (instead of data for individual holdings). Abnormal observations are then traced back to the holdings and treated at that level.

Like so many other branches of statistics in different countries, Swedish agricultural statistics have experienced an increase in nonresponse rates in recent years. That tendency is certainly unhappy, though in most cases the rate is still comparatively low. To exemplify, the nonresponse is about 0.2 % in the Farm Register and about 2 % in the Objective Crop Yield Surveys, but is as large as about 30 % in the Farm Bookkeeping Survey.

The ambition to develop new statistics on changes has resulted in some severe quality problems. Measurement errors of small or even negligible importance in traditional statistics may indeed be significant if the data are used in a longitudinal study. The proper handling of these problems will certainly call for prolonged work.

4.4. Timeliness

It is easy to state that the value of a statistic is highly dependent on when it appears. However, experience from agricultural statistics in Sweden has shown that the situation is in reality somewhat more complex. The users of the statistics tend to adapt their procedures and routines to when the statistics have become available in the past. As a consequence it may take some time before the real effects of improved timeliness, both up-to-dateness and punctuality, become apparent. Even if the degree of timeliness has not been criticized much, a special drive for improvement has been regarded as an essential part of the development programme. As a result some statistics are now released considerably earlier than before. For a general discussion with examples from Swedish agricultural statistics, see Medin (1984).

4.5. Effects of developments in general statistical methodology and production techniques

The general methodological and technical progress in official statistics during the postwar period has, of course, been a great impetus for the development of Swedish agricultural statistics. First, the basic goal of using modern sampling in Swedish agricultural statistics was to avoid or reduce biases. Then questions about precision were discussed. Better sampling procedures have also been regarded as a valuable rationalization measure, especially given the financial stringency of recent years. For the history of sampling in Swedish agricultural statistics, see Dalenius (1957) and Medin (1983).

At Statistics Sweden the Objective Crop Yield Surveys were among the very first to apply electronic data processing (EDP). When powerful computers became available automated routines were developed for the annual surveys on crop areas and livestock numbers as described by Medin and Larson (1964) and Wilson (1967). Developments in agricultural statistics later on were closely linked to the remarkable developments in statistical data processing in general. (Nevertheless, against the background of their early experiences in agricultural statistics the authors of this article cannot refrain from wondering why the new technique has not had an even more profound effect on official statistics with regard to the specific statistical aspects of the work. This includes for instance data editing and the development of entirely new kinds of statistical results.)

In Sweden register techniques have been used for official statistics ever since the first census of population in 1749. However, the role of registers in official agricultural statistics has been limited until recent decades, during which two lines of development have emerged. The one is the use of administrative records; and an example is the regular use of

tax assessment data as a basis for farm income statistics. The other, which is by far the most important, dates back twenty years to the establishment of the Farm Register. Organized as a total panel, the register has opened up new means for agricultural statistics and their coordination.

For years, the extent to which central statistical offices ought to engage in statistical analysis has been the subject of debate. This was demonstrated for instance at the Washington seminar in 1977 on "Statistical services in ten years' time" arranged by the Conference of European Statisticians, see Duncan (1978). However, in recent years at Statistics Sweden there has been a clear tendency towards more analytical activities. In agricultural statistics this subject was treated by a special working party. One conclusion drawn was that analytical activities should not be separated from the more traditional statistical activities. On the contrary, to fulfil analytical aims it is often necessary to let these aims influence basic statistical procedures such as sample design and data collection. A list of potential projects was presented; many of these are now being implemented. An example, not mentioned elsewhere in this article, is the systematic application of multivariate statistical techniques to the data from the Objective Crop Yield Surveys to look for factors of importance for the level of yield obtained under regular farming conditions. Certain preliminary results have been presented by Statistics Sweden (1985 – 86).

5. Comments on Budgetary and Other Restraints

As is so often the case in official statistics, development work has been dependent on overcoming a number of hindrances. Some of these hindrances have in fact stimulated improvements in methods and routines.

Respondent burden often represents an important restriction. In Swedish agricultural statistics, proposals for new data collections have always been treated in a careful manner. However, during the latter part of the 1970s, there was a considerable public debate about the responsibility of enterprisers to answer statistical questionnaires. For that reason a special study of the time spent by farmers on reporting agricultural statistics was initiated. A rough estimate was that, on average, the farmers spent about one hour a year on this task, but a small group of farmers participating in several sample surveys spent considerably more time, see Hedqvist and Rösiö (1984). The conclusion drawn was that, generally speaking, farmers are not overburdened. Consequently, no extra restraint was imposed.

Other kinds of restrictions follow from the legislation relating to confidentiality and integrity in connection with data collection and data storage. Those questions have been much debated in Sweden in recent years, not least with reference to official statistics. A Data Privacy Protection Act was passed in 1974. According to this act permission by the Data Inspection Board is often required in advance of a data collection for statistical purposes. However, the existence of legal provisions for confidentiality and integrity can increase the readiness of respondents to provide correct information. The Farm Register represents a special case. Although kept by Statistics Sweden it is used for both statistical and administrative purposes. This has called for and resulted in a separate government statute for the register but the twofold use of the data has in practice not been the cause of any great difficulty.

Due to existing budgetary restraints the total cost for statistics has not been allowed to increase in recent years; in fact total cost has had to decrease because of financial restraints prescribed since 1978. The importance of this cost restriction is difficult to interpret. Improvements in official statistics

can very well be achieved with reduced funding. A basic reason is that prerequisites for statistical work are changing all the time. Proper adjustments to these changes make new resources available; for more detailed arguments see Medin (1984). However, for the sake of completeness it should be added that in Swedish agricultural statistics, limited cost reductions have also been achieved in recent years by direct cuts in the production programmes: eliminated variables, reduced regional breakdown, smaller samples, etc.

6. Development Projects: Some Methodological Issues

6.1. *Classification by type of farming*

During recent decades farm typologies, i.e., classifications of farms according to type of farming, have been introduced in official agricultural statistics in a large number of countries. In Sweden work on a farm typology started in 1977, when a committee on farm typology was appointed with representatives for Statistics Sweden and principal users of agricultural statistics. The formal decision to introduce the new classification was taken in 1982.

The typology is used to classify the farms into a limited number of groups according to their production patterns. In the Swedish typology this is achieved by combining data on crop acreages and livestock numbers with standard labour requirements per hectare and per animal. Depending on the relative sizes of the total labour requirements in various enterprises, each farm is classified as a grain farm, a dairy farm, a pig farm, etc.

The classification is closely linked to the Farm Register which contains the acreage and livestock data needed. For each farm, its type is recorded annually in the register and from there transferred to almost all areas of agricultural statistics. Thus, structural statistics can now convey information on the specialization in farming, while the statistics on farm economics contain particulars on how profitability varies between type groups.

The experience gained so far – including the very good reactions from users of agricultural statistics – clearly indicates that the farm typology is a valuable enhancement. Better insight into actual farming conditions has been achieved. In this connection it is worth noting that the costs involved are very small indeed since all necessary data already exist. For details see Wilson (1974), Jorner (1979), Medin (1985), Typologigruppen (1979, 1982), and Statistics Sweden (1987a).

6.2. *Farm models*

Based on ideas applied in Norway and Finland the concept of "farm models" has recently been introduced in Swedish discussions about the data basis for agricultural policy decisions. The idea is to "construct" a small set of farms with specialized and well-defined production patterns and to make synthetic calculations of production costs, revenue, income, etc. for these farms. These calculations – which should be based on accepted principles of farm business economics – will use agricultural statistics, standard values, forecasts, etc.

Compared with ordinary statistics the results should represent farms with more clearly defined production patterns. Another objective is to produce data for periods for which statistics are not yet available, e.g., present or future years.

An interagency working group has been established to develop and test the idea of farm models; Statistics Sweden is represented in that group. For the work done in Norway and Finland, see Budsjettnemnda for jordbruket (1986) and Ikonen (1985), respectively.

6.3. Longitudinal analysis

On the basis of the Farm Register, longitudinal studies on structural changes in farming have been carried out annually since 1971. To make this possible, the traditional definition of the holding had to be supplemented with rules referring to time. These rules answer questions such as: "If two farms have been amalgamated, which of them, if any, shall be regarded as remaining?" or "If a farm has been divided into two, which of these, if any, shall be regarded as a continuation of the original?" With the existence over time of the register units thus established, it has become possible to observe changes from year to year in a farm's size and type. These statistics are presented in matrix tables which show the transitions between different classes of holdings, for instance size classes of arable land or size classes of herds. These tables also contain data on the number of holdings that have started up or fallen into disuse during the year, see Medin and Wilson (1974, 1985) and Statistics Sweden (1987b). Future development work in this field is intended to include other areas of animal husbandry statistics and also forestry statistics.

6.4. Projections and forecasts

To aid decision-making by political bodies as well as in administration and business, agricultural statistics often have to be supplemented by projections and forecasts. These can, however, in many cases be said to predict future statistics and consequently the statistical agencies have to face the question of what their role should be. In the review of Swedish agricultural statistics the conclusion was that new projections and forecasts will probably be short-term and in fields where Statistics Sweden is already responsible for the corresponding statistics.

At this point, reference should be made to development work on objective crop yield forecasts which was carried out in 1966 – 1975, see Statistics Sweden (1976a). The idea was to base the forecasts on field measurements of straw length, number of ears, etc., but also on meteorological information. The results were fairly encouraging, apart from the fact that a regular forecasting service would have been rather costly. It could therefore not be realized. Some years later Statistics Sweden cooperated in developing a model for crop yield forecasts based solely on meteorological data. Such forecasts are now made by the National Agricultural Market Board; see Rösiö, Tillgren, and Loman (1979).

To the present, most of the forecasting work within agricultural statistics at Statistics Sweden refers to changes in the agricultural structure. A method linked to the above-mentioned longitudinal statistics received from the Farm Register has been developed. Transition matrices are used to compute Markovian projections over one to fifteen years, see Thorburn (1980, 1981, 1983), Wilson and Jorner (1984) and Statistics Sweden (1985b).

6.5. Use of geographical coordinates

Starting with a paper by Hägerstrand (1955) extensive development work has been in progress in Sweden on the use of geographical coordinates. The basic idea is that registers of real estates, houses, work places, archaeological finds, etc., should be supplemented with information on their latitude and longitude. For point objects this is fairly straightforward but for line objects such as roads or boundaries and area objects such as fields or farms, it is usually necessary to make approximations.

For official statistics the availability of geographical coordinates creates new opportunities; some examples from agricultural statistics follow.

Traditionally, statistical objects such as farms are classified according to pre-fixed divisions into parishes, local authority areas, counties, natural farming areas, etc. If coordinates are available, statistics can easily be produced for any required subdivision by introducing the coordinates of the new geographic boundaries. Suppose, for instance, that the geographical boundaries between different types of soil can be entered into a computer. If coordinates were available for all sample plots in the Objective Crop Yield Surveys, yield statistics for the various soil type districts could easily be calculated. Today such recalculations involve costly procedures.

It seems reasonable to assume that geographical coordinates will, in the future, create many new types of statistics. This may include regional ex post classifications, studies of distances, for instance, between farm centre and farm fields or between farm centre and different commercial centres, new routines for editing primary data, or new techniques for presenting agricultural statistics in the form of maps.

As to current developments in Sweden, geographical coordinates are being systematically recorded in the new EDP-based system for official land registration. This information will probably soon be available for use in some areas of agricultural statistics.

6.6. Spatial analysis

The rapid developments in the analysis of spatially-defined data during the last few years are well documented in the literature; for references see, e.g., Wilson and Bennett (1985) or Upton and Fingleton (1985). Obviously, many aspects of agriculture naturally lend themselves to this type of analysis. Therefore, in connection with the development work presented in this article a special working party was set up in 1983 to study spatial reporting in among others, agricultural statistics. The result of the work has been published by Statistics Sweden (1984b). The following are a few of the many ideas presented in that report.

1. In agricultural statistics the regional breakdowns have traditionally been decided on in advance of the tabulations. An alternative approach would be to use the data collected as a basis for estimating isarithms, distinguishing different categories of land. A simple example could be the use of data from the Objective Crop Yield Surveys to divide the country into high yield and low yield districts (or according to some more developed yield level classification).
2. In sample surveys the material available is often too small to allow for a very detailed geographical breakdown. By use of models one could perhaps estimate average yields or total use of fertilizers in small areas such as municipalities.
3. For many subject matter areas, closer studies into the geographical variation would be appropriate. Areas mentioned include crop varieties, type of farming, water availability, etc.

7. Development Projects: Some Subject Matter Issues

In the present section the reader will find brief descriptions of fields where certain projects have been proposed. Most of these projects refer to variables which have so far not been included in Swedish agricultural statistics.

7.1. Agricultural structure

Fundamentally, the purpose of statistics on agricultural structure is to show the properties of the farms and the state and changes of its production factors: labour force (including the holder and his or her family), land and

other real capital (buildings, machinery, livestock, etc).

Most of these statistics are based on the Farm Register. The annual statistical reports thus contain regionally differentiated data on both the state of and the (net and gross) changes in the total population of farms. Special studies regarding the structure of both crop and animal husbandry have been carried out as parts of the 1981 Census of Agriculture as reported by Statistics Sweden (1984a). In this connection reference may also be made to Medin and Wilson (1985).

Development work in the statistics on agricultural structure has to be continued while new political and economic aspects need to be analyzed. For instance, one important aim is to highlight the continuing specialization process in crop or animal production. Resources should also be used to illustrate in depth the changes in ownership and tenancy, and to what extent holdings are engaged in both agriculture and forestry.

7.2. *Family farming*

The term "family farming" has long been accepted in farm economics research, see, e.g., Warren (1920), Scoville (1947) and Nordiska Jordbruksforskares Förening (1983). A related concept in the German language is "Bäuerlicher Landwirtschaft," see Neander (1983). Terms like these also have a central position in the agricultural policy debate in Sweden. However, as pointed out by Swedborg (1980), it is noteworthy that the exact meaning of the terms has never been clarified. Even if such a clarification is not immediately necessary, the frequent references to the concepts may indicate a potential need for new statistics. A development project has therefore been proposed.

7.3. *Part-time farming*

Despite the political efforts in the post-war period to promote full-time farming, part-time farming has become much more common. As in other Western countries, see Martens (1980), what is called "dual job holdings" or "multiple job holdings" are now very frequent all over Sweden.

However, from the statistical point of view, part-time farming has not been studied to any great extent in Sweden. For that reason Statistics Sweden now cooperates with other government authorities in a special project. The aim is to develop and systematize concepts, definitions, and classifications for part-time farms as well as part-time farmers, and to describe the structure and the importance of part-time farming.

7.4. *Agricultural infrastructure*

Modern farms are highly dependent on their infrastructure, by which we mean "the financial, institutional, and social surroundings which influence the development, survival, and production of the farms." That tentative definition was presented by a special working party appointed to study how the agricultural infrastructure can be illuminated by agricultural statistics.

As for variables, the statistics may, for instance, refer to the availability of (for the farm as such and for the people living there) communications, slaughterhouses, dairies, machine stations, distributors of seed and fertilizers, shops, schools, medical care, etc. Variables related to the labour market, such as manpower supply on one hand and off-farm job opportunities on the other, may also be included.

No decision has yet been taken to develop regular statistics on the agricultural infrastructure. Meanwhile, a compendium containing available official and other statistics will be issued.

8. Development Projects: Short Notes on Some Additional Topics

Although new methods, new techniques, and new subject matter areas are very important to the development of official statistics, most progress is certainly a result of new ways to combine procedures used before. For the sake of balance some projects of this character will be briefly mentioned. In addition some methodological studies not presented elsewhere will be named.

New types of labour force statistics for agriculture based on telephone interviews has been introduced, see Hedqvist (1982).

Postal surveys on farmers' assets and liabilities have been carried out.

An ad hoc study of field irrigation in agriculture has been performed in two steps. First, data on the availability of irrigation installations were collected by the Farm Register. Then, from those farmers who reported such installations, details regarding irrigated crop areas, irrigation techniques, types of water resources, etc., were collected by mail. It may be added that this way of collecting statistical data, or variations of it, has over the years been used on a number of occasions for studying "rare items." An alternative approach to the study of such items in agriculture has recently been presented in this journal by Fesco, Tortora, and Vogel (1986).

Horticultural statistics have been reorganized and expanded. Among other things this has meant that Censuses of Horticulture are now regularly taken at three-year intervals. This is a higher frequency than before.

In Sweden, animal husbandry statistics have traditionally been very limited, at least in comparison with crop husbandry statistics. This is so in spite of the fact that most of the farm income normally stems from meat and milk production. Also in the development work presented here only limited proposals have been made; they refer, for instance, to the present and future structure of animal husbandry.

New statistics on taxes and other transfers (family allowances, housing grants, etc.) have been compiled in connection with the Survey of Farmers' Assessed Incomes, Expenditures, Net Earnings, etc.

For the period up to 1955 historical agricultural statistics have been published in a separate volume by Statistics Sweden (1959). In the preparatory work for an updated edition, special emphasis will be placed on the quality of the statistics, in particular with regard to comparability over time. Problems involved have been discussed by Ribe (1982).

Special methodological studies have been made regarding existing or potential statistics on:

1. owner and holder relationships,
2. tenancy conditions,
3. building construction,
4. gross changes in the area of arable land (i.e., in areas reclaimed or taken out of use).

9. Final Comments

In the history of Swedish agricultural statistics, major changes have been caused by, e.g., new statistical needs, new economy measures, new organizational structures, or – rather often – the introduction of new statistical methods and techniques. The work presented in this paper does not have a monolithic background but can be characterized by the ambition to progress through a large number of limited efforts (made with due regard to existing financial restraints).

10. References

Most of the references given here have been included because of their methodological content. This also applies to the official statistical reports referenced.

Tabular results from the statistical work decribed in this article can, in most cases, be found either in the Yearbook of Agricultural Statistics or in the Statistical Reports series ("Statistiska meddelanden" with a sub-series for agricultural statistics). Both are published by Statistics Sweden.

A bibliography on Swedish agricultural statistics 1950 – 1974, including tabular presentations as well as methodological descriptions and studies, was published by Statistics Sweden (1975).

Borglund, D., Jorner, U., Medin, K., Olofsson, P.O., and Polfeldt, T. (1984): Sifferfabrik eller statistikverk – det statistiska synsättets betydelse för statistiska centralbyrån. (Figures factory or statistical office – the importance of the statistical approach for Statistics Sweden. In Swedish.) Statistisk tidskrift (Statistical Review), 22 (3), pp. 171 – 176.

Budsjettnemnda for jordbruket (1986): Modellbruksberegninger. Regnskapstall for 1984. Fremregnede tall for 1985 og 1986. (Calculations for farm models. Estimates for 1984. Projections for 1985 and 1986. In Norwegian.) Mimeo.

Dalenius, T. (1957): Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice. (Diss. Uppsala.) Stockholm/Uppsala.

Duncan, J.W. (Ed.) (1978): Statistical Services in Ten Years' Time. Pergamon Press.

Fesco, R., Tortora, R.D., and Vogel, F.A. (1986): Sampling Frames for Agriculture in the United States. Journal of Official Statistics, 2 (3), pp. 279 – 292.

Hägerstrand, T. (1955): Census Returns, Air Photographs and Data-processing Machines. A project for Combination. (In Swedish with summary in English.) Svensk geografisk årsbok 31, pp. 233 – 255.

Hedqvist, L. (1982): Labour in Agriculture – A New SCB Survey. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 20 (3), pp. 157 – 172, 232 – 234.

Hedqvist, L. and Rösiö, G. (1984): Time Spent on Reporting Data to Agricultural Statistics. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 22 (4), pp. 339 – 344, 399 – 400.

Hedqvist, L. and Rösiö, G. (1985): Development Work in the Agricultural Statistics of Statistics Sweden. (In Swedish with summary in English.) Jordbruksekonomiska meddelanden 1985:9, pp. 336, 338 – 346.

Hedqvist, L. and Thorburn, D. (1978): Development Trends in Agricultural Statistics from the Central Bureau of Statistics. (In Swedish with summary in English.) Jordbruksekonomiska meddelanden 1978:7 – 8, pp. 206 – 207, 208 – 217.

Ikonen, J. (1985): Use of Calculations of Production Costs and Bookkeeping Results in the Follow-up of Farmer's Incomes. Paper presented to Finnish-Hungarian-Polish seminar, November 1985.

Jönrup, H. (1976): Sampling and Estimation Processes of the Objective Crop Yield Estimation in Sweden. Statistisk tidskrift (Statistical Review), 14 (5), pp. 402 – 412

Jorner, U. (1979): Type Classification of Agricultural Enterprises – A Way of Extracting More Information from Collected Data. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 17 (6),pp. 439 – 452, 483 – 484.

Larsson, G., Medin, K., and Wilson, B. (1987): A Farm Bookkeeping Survey as Part of Official Agricultural Statistics: The Case of Sweden. Statistical Journal of the United Nations Economic Commission for Europe, 4 (3), pp. 245 – 257.

Martens, L. (1980): Part-time Farming in Developed Countries. European Review of Agricultural Economics, 7 (4), pp. 377 – 393.

Medin, K. (1965): Crop Yield Estimation and Crop Insurance in Sweden. Review of

the International Statistical Institute, 33(3), pp. 414 – 442.

Medin, K. (1983): The Introduction of Probability Sampling in Swedish Acreage and Livestock Statistics in 1950. A Note on the Background and Later Developments. Essays in Honour of Tore E. Dalenius. Statistisk tidskrift (Statistical Review), 21 (5), pp. 19 – 25.

Medin, K. (1984): Timeliness in the Production of Official Statistics. Statistisk tidskrift (Statistical Review), 22 (1), pp. 5 – 15.

Medin, K. (1985): The Farm Register Approach in Sweden – Principles and Potentialities. Bulletin of the International Statistical Institute, Vol. LI, Book 2, pp. 13.2.1 – 15.

Medin, K. and Larson, B. (1964): The New System for Agricultural Statistics. II. Automatic Control and Correction of Primary Data. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 2 (6), pp. 393 – 403, 466 – 467.

Medin, K. and Wilson, B. (1974): Farm Structure in Figures. A Study in Statistical Methodology. European Review of Agricultural Economics, 1(4), pp. 461 – 481.

Medin, K. and Wilson, B. (1985): Measuring Changes in the Size and Type of Farms – Some Swedish Data. Paper presented to poster session at the XIX International Conference of Agricultural Economists in Malaga 1985 arranged by the International Association of Agricultural Economists.

Neander, E. (1983): Zur Abgrenzung, Charakterisierung und Bewertung bäuerlicher Landwirtschaft. Berichte über Landwirtschaft. Zeitschrift für Agrarpolitik und Landwirtschaft. Herausgegeben vom Bundesministerium für Ernährung, Landwirtschaft und Forsten, 61(1), pp. 67 – 78. (In German)

Nordiska Jordbruksforskares Förening (1983): Familjelantbrukets framtid. (Scandinavian Association of Agricultural Scientists: Future of the family farm. In Scandinavian languages.) NJF-utredning/rapport nr 11. (See also Nordisk jordbruksforskning 1983:3, pp. 454 – 462 and 1983:5, pp. 791 – 792.)

Polfeldt, T. (1977): Reexamination of the Area Checking Surveys. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 15 (3), pp. 202 – 218, 269 – 271.

Ribe, M. (1982): On Republication of Old Agricultural Statistics. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 20 (1), pp. 5 – 23, 61 – 62.

Rösiö, G., Tillgren, U., and Loman, J. – O. (1979): Crop Forecasts Based on Weather Observations. (In Swedish with summary in English.) Jordbruksekonomiska meddelanden 1979:1, pp. 8 – 9, 10 – 24.

Scoville, O.J. (1947): Measuring the Family Farm. Journal of Farm Economics, Vol. XXIX, pp. 506 – 519.

Statistics Sweden (1936): Le recensement général agricole de 1932. (In Swedish with summary in French.).

Statistics Sweden (1959): Historical Statistics of Sweden II. Climate, Land Surveying, Agriculture, Forestry, Fisheries – 1955. (In Swedish with foreword, table headings, etc. in English.) Stockholm.

Statistics Sweden (1975): Bibliografi över den svenska jordbruksstatistiken 1950 – 1974. (Bibliography of Swedish Agricultural Statistics 1950 – 1974. In Swedish.) Promemorior från SCB, 1975:8.

Statistics Sweden (1976a): Objektiva skördeprognoser. Utredningsrapport. (Objective Crop Yield Forecasts. Report. In Swedish.).

Statistics Sweden (1976b): Komparativ studie av jordbruksstatistiken i de nordiska länderna och övriga Västeuropa. (Comparative Study of Agricultural Statistics in the Nordic Countries and the Rest of Western Europe. In Swedish.) Mimeo.

Statistics Sweden (1978): Utvecklingslinjer i SCBs lantbruksstatistik. (Development Trends in Agricultural Statistics from Statistics Sweden. In Swedish.) Mimeo.

Statistics Sweden (1983): Information about the National Swedish Farm Register. Mimeo.

Statistics Sweden (1984a): The 1981 Agriculture Census. Special Studies into Type of Farming and Structure in Cropping. (In Swedish with summary in English.) Statistiska meddelanden J 1984:18.

Statistics Sweden (1984b): Den rumsliga redovisningen i areell statistik. Idépromemoria framtagen av en särskild arbetsgrupp. (Spatial Reporting in Area Statistics. Memorandum on ideas presented by a special Working Party. In Swedish.) Mimeo.

Statistics Sweden (1985a): Utvecklingsinsatser inom SCBs lantbruksstatistik. (Development Efforts in Agricultural Statistics at Statistics Sweden. In Swedish.) Mimeo.

Statistics Sweden (1985b): Projections of Structural Changes up to the Year 2000. Number of Holdings by Size Group, Type of Holding and Region. (In Swedish with summary in English.) Statistiska meddelanden J 30 SM 8504.

Statistics Sweden (1985–86): Analys av skördevariationer – med utgångspunkt i de objektiva skördeuppskattningarna. Delrapporter 1 – 2 från förstudie. (Analyses of Crop Yield Variations – Based on the Objective Crop Yield Surveys. Progress reports 1 – 2 from preliminary study. In Swedish.)

Statistics Sweden (1987a): Type of Farming on the 13th of June, 1985. (In Swedish with summary in English.) Statistiska meddelanden J 30 SM 8603.

Statistics Sweden (1987b): Structural Changes in Agriculture 1985 – 1986. Longitudinal Data. (In Swedish with summary in English.) Statistiska meddelanden J 30 SM 8702.

Statistics Sweden (1987c): Objective Crop-Yield Surveys in Sweden. Memoranda Series: 1987: 1.

Statistics Sweden (1988): The Swedish Crop Insurance System. Principles and Methods. Mimeo.

Swedborg, E. (1980): Lantbrukspolitik för 80-talet. (Agricultural Policy for the 1980s. In Swedish.) Stockholm.

Söderlind, T. (1982): Development of the Swedish Objective Crop Yield Surveys. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 20 (2), pp. 103 – 121, 143 – 145.

Thorburn, D. (1980): Forecasting the Agricultural Structure Using Empirical Transition Matrices. European Review of Agricultural Economics, 7 (4), pp. 413 – 432.

Thorburn, D. (1981): Projections by Means of Longitudinal Studies and Transition Matrices. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 19 (3), pp. 189 – 201, 238 – 239.

Thorburn, D. (1983): Forecasting Aggregate Time Series Using Empirical Transition Matrices. Scandinavian Journal of Statistics, 10, pp. 35 – 39.

Typologigruppen (1979): Klassificering av de svenska jordbruksföretagen efter driftsinriktning och driftens omfattning. (Committee on Farm Typology: Classification of Swedish Farms by Type of Farming and Size of Business. In Swedish.) Lantbruksekonomiska samarbetsnämndens sektorsgrupp. Mimeo.

Typologigruppen (1982): Typklassificering av jordbruksföretag. Resultat och slutsatser från försöksverksamheten. (Committee on Farm Typology: Type Classification of Farms. Results and Conclusions from the Test Studies. In Swedish.) Lantbruksekonomiska samarbetsnämndens sektorsgrupp. Mimeo.

Upton, G.J.G. and Fingleton, B. (1985): Spatial Data Analysis by Example. Vol. 1. John Wiley & Sons.

Warren, G.F. (1920): Farm Management. The Macmillan Company, New York.

Widén, M. – L. and Åstrand, H. (1975): Agricultural Statistics of the Future. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 13 (3), pp. 224 – 243, 268.

Wilson, A.G. and Bennett, R.J. (1985): Mathematical Methods in Human Geography and Planning. John Wiley & Sons.

Wilson, B. (1967): The New System for Agricultural Statistics. V. Manual Routines. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 5 (1), pp. 36 – 46, 64 – 66.

Wilson, B. (1974): Classification of Agricultural Enterprises. (In Swedish with summary in English.) Statistisk tidskrift (Statistical Review), 12 (2), pp. 145 – 153, 179 – 181.

Wilson, B. and Jorner, U. (1984): Strukturprognoser inom jordbruket – praktiska och teoretiska aspekter. (Forecasting the Agricultural Structure – Practical and Theoretical Aspects. In Swedish.) Report from the research conference on longitudinal studies arranged by Statistics Sweden in Norberg, October 1984, pp. 77 – 80.

Zarkovich, S.S. (1977): Sample Surveys for Area and Yield Statistics. Bulletin of the International Statistical Institute, Vol. XLVII, Book 3, pp. 440 – 452. (With discussion pp. 453 – 461).

Received June 1987
Revised December 1987

Journal of Official Statistics
Vol. 4, No. 2, 1988, pp. 191–199

Book Reviews

Books for review are to be sent to the Book Review Editor Jan Wretman, Statistical Research Unit, Statistics Sweden, S–115 81 Stockholm, Sweden.

Gelfand, A.E. (Ed.), Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon. Academic Press, Inc., Boston, 1987, ISBN 0-12-279450-8, xxviii + 544 pp., $ 59.95.

In 1984, Herbert Solomon celebrated his 65th birthday and his 25th year at Stanford University. To his honor, this volume of contributed papers was prepared and published. Like most books of this type, the contents are heterogeneous. This also reflects the diversity of Solomon's own interests and contributions, including, for instance, operations research, geometrical probability, multivariate analysis, and "jurimetrics." After a biographical sketch and a list of Solomon's publications, the book provides twenty papers grouped into four sections. The contents are well summarized by the following list of informative titles.

1. Operations Research and Applied Probability

Inequalities for Distributions With Increasing Failure Rate (M. Brown).
A Markov Decision Approach to Nuclear Materials Safeguards (H. Chernoff & Y. Yao).
On the Persistent Release of Particles in a Fluid Flow (J. Gani & P. Todorovic).
Statistical Inference for Random Parameter Markov Population Process Models (D. Gaver & J.P. Lehoczky).

2. Distribution Theory and Geometric Probability

Probabilistic-Geometric Theorems Arising From the Analysis of Contingency Tables (P. Diaconis & B. Efron).
Some Remarks on Exchangeable Normal Variables With Applications (S. Geisser).
Asymptotics for the Ratio of Multiple t-Densities (S.J. Press & A.W. Davis).
Periodogram Testing Based on Spacings (A.F. Siegel & J. Beirlant).
Tests for Uniformity Arising From a Series of Events (M.A. Stephens).
Spatial Classification Error Rates Related to Pixel Size (P. Switzer & A. Venetoulias).

3. Applications

The Use of Peremptory Challenges in Jury Selection (M.H. DeGroot).
An Information-Processing Model Based on Reaction Times in Solving Linear Equations (J.B. Kadane, J.H. Larkin & R.E. Mayer).
Diagnostic Errors and Their Impact on Disease Trends (M.A. Kastenbaum).

Hypothesis Testing in the Courtroom (D.H. Kaye).
Multivariate Discrimination of Depressive Groups Across Cultures (J.E. Mezzich & E.S. Raab).

4. Inference Methodology

Estimation in Parametric Mixture Families (A.E. Gelfand).
Multiple Shrinkage Generalizations of the James-Stein Estimator (E.I. George).
The Analysis of a Set of Multidimensional Contingency Tables Using Log-Linear Models, Latent-Class Models, and Correlation Models: The Solomon Data Revisited (L.A. Goodman).
Selection Procedure for Multinomial Populations with Respect to Diversity Indices (M.H. Rizvi, K. Alam & K.M.L. Saxena).
Confidence Intervals for the Common Variance of Equicorrelated Normal Random Variables (S. Zacks & P.F. Ramig).

Gani and Todorovic discuss interesting applications of probability, although the paper is of less importance than previous work on the same topic by the authors.

Gaver and Lehoczky study transitions in a population between "colonies" or "compartments," where each colony has an unknown migration rate parameter. The likelihood for a "simple Markov population process" (SMPP) is easily seen to be of multivariate independent Poisson type. The authors also allow heterogeneity by assuming that several realizations of an SMPP are observed, corresponding to rate vectors independently drawn from a superpopulation. First, Bayesian posterior distributions are calculated for the realized rate parameters under a few completely specified superpopulations. Next, a bit more realistic, these superpopulations are allowed to depend on unknown parameters, simultaneously estimated by empirical Bayes methods. Formulas are derived, but the paper does not provide much insight.

Diaconis and Efron show relations between some classical probabilistic theorems. This is an entertaining paper in the spirit of Feller's books.

Stephens contributes a discussion of tests for the Poisson process property of a series of events. He reviews and compares conventional and special tests against various types of alternatives. The paper appears to be a valuable addition to the literature on this subject.

Kastenbaum discusses data quality in medical diagnoses and death certificates, considering, in particular, cancers in the United States. Misclassifications and varying coding practice affect official vital statistics and can make conclusions about disease or mortality rates questionable. The rate of autopsy confirmation in the United States has decreased, which renders assessment of error frequencies even more difficult. As his main example, Kastenbaum discusses the differences between the 1968 and 1977 U.S. lung cancer death rates. Of particular value is the concluding chronological bibliography on the subject, containing more than a hundred references from 1955 to 1984.

The contribution by Gelfand concerns admissibility, Bayes estimation, and empirical Bayes estimation in parametrically specified mixtures of a distribution family. A typical example of the paper is obtained by scaling a chi-square distribution by a scaling factor that has a Poisson mixing distribution with an unknown intensity parameter. The paper is unconvincing about the method's potential for applications.

The paper by Goodman is the longest paper in this collection. Starting from a 2^5 cross-classification data set analyzed by Solomon in 1961, Goodman reviews some models for categorical data, of log-linear, linear, and latent-class type. The presentation is elementary and in some respects detailed. It provides references to previous papers by the author rather than new ideas.

Behind this collection is a distinguished group of authors. However, the heterogeneity of subjects in combination with a specialized and technical character makes the book a natural purchase only for statistical libraries. Several contributions are reports of the current status of ongoing research and are likely soon to be outdated by more comprehensive or conclusive papers by the same authors in other publications. Unfortunately, the text is only of typewriter quality.

Rolf Sundberg
University of Stockholm
Stockholm
Sweden

Kish, L., Statistical Design for Research. John Wiley & Sons, Inc., New York, 1987, ISBN 0-471-08359-3, xxii + 267 pp., $ 34.95.

This book provides an excellent description of issues that must be considered in developing or evaluating designs for empirical investigations. Three categories of designs are assessed, namely, experimental designs, sample surveys, and observational studies. The material is kept nontechnical and concise with abundant references to other sources for technical details. This book can be read by any researcher regardless of his or her statistical training.

In Chapter 1, criteria are developed to compare design options based upon what should be the 3 R's of good design: Representation, Randomization, and Realism. These criteria aim at creating the design closest to the ideal, i.e., one that facilitates unbiased estimation and testing. The three basic design categories are described in relation to this ideal. I particularly appreciated Kish's distinction between the goals and aims of survey sampling and those of experimental design.

Chapter 2 expands on these concepts in describing the design features of survey samples. Four population levels are defined and the effects of nonresponse and undercoverage are distinguished for each level. Key to sample design is the use of design effects. Kish presents a thorough description of this concept for estimates based upon both the total sample and on subclasses. Then, design characteristics such as clustering and stratification are described and their impact on the design effect summarized.

Observational studies are often used when probability sampling is impractical. Five basic designs for observational studies are presented in Chapter 3. Models are developed to contrast these designs in terms of variance and bias for a fixed total cost. Four major types of bias are distinguished and the ability of each design to reduce or eliminate each type of bias is discussed. While I liked the model-based comparison of the designs, I thought the presentation was not as straightforward as it could have been nor was it as easy to read as the rest of the book. Total cost is set to what would be achieved by the one-shot case study. Thus, other designs include sample size reduction factors in their variance expression to reflect the additional costs associated with design features such as control groups, pretests, etc. Many of the studies that I have been involved with have demanded fixed precision instead of cost. To use the material presented here would require reworking the variance expression and developing an explicit cost model.

Unfortunately, the author does not include a review of the conceptual principles underlying experimental designs similar to the presentations in Chapters 2 and 3 for surveys and observational studies. While the mathematical underpinnings of experimental designs are treated in numerous texts, a discussion of the thought processes that lead to choosing one design over another is often lacking in these texts. Kish's book would have benefitted from such a discussion, using the same unified approach as he uses elsewhere in his book.

Extraneous variables can disguise treatment effects or population characteristics of interest. Chapter 4 provides an excellent summary of methods for controlling the biasing or variance-inflating effects of extraneous variables. These methods include design features such as stratification in surveys, or blocking in experimental designs, and analysis procedures such as post-stratification and standardization. Guidelines are presented for choosing the best set of control variables and for deciding whether to control by design or by analysis.

Samples and censuses are contrasted in Chapter 5. Various spinoffs of these designs are discussed, for instance, registers and samples associated with censuses. The presentation is not adequately specific, which limits the usefulness of this chapter. For instance, small area estimation is alluded to but not fully addressed.

Sample designs measuring population characteristics that vary with time is the topic of Chapter 6. The characteristics of panel, repeated, periodic and overlapping surveys are described in terms of variance reduction and bias control. The presentation is detailed and, to my knowledge, not readily available elsewhere.

Chapter 7 consists of miscellaneous material the author thought too technical to present

elsewhere. There are some good nuggets of information that make the chapter worthwhile. For instance, the practical requirements for a measurable survey design are summarized and strategies for satisfying multiple survey objectives are explained.

Overall, I found this book highly readable. It presents concepts of design that researchers have had to discover by trial and error or through discussions with more experienced colleagues. I believe this book would be valuable reading for practicing statisticians. Professors teaching classes in experimental design, sampling, or social research are encouraged to review these concepts with their students in addition to materials traditionally presented. In short, this book may not be perfect but it is very good and I recommend it.

Brenda G. Cox
Research Triangle Institute
Research Triangle Park, NC
U.S.A.

Little, R.J.A. and Rubin, D.B., Statistical Analysis with Missing Data. John Wiley & Sons, Inc., New York, 1987, ISBN 0-471-80254-9, xiv + 278 pp., $ 34.95.

Standard methods for the analysis of statistical data usually assume that all intended observations are available. Hence, the observations can be stored in a rectangular matrix, the rows representing records or cases, and the columns variables. However, for various reasons values may be missing in this matrix; respondents in a sample survey may refuse to answer some or all of the questions, or controlled experiments may break down due to mechanical failures.

The aim of the book is to survey current methods for the treatment of missing data. Particularly in survey sampling, missing data caused by nonresponse is a serious problem. Since nonresponse rates tend to increase, the problem becomes more and more serious. Therefore, this book is important for those who are involved in both theoretical research and practical applications of survey sampling.

The first part of the book covers the history of the missing data problem in three important areas of statistics: analysis of variance, survey sampling, and multivariate analysis. The introductory chapter starts with a general overview. One approach is to discard all incomplete records and use only complete records in the analysis. For a small number of missing values this approach may work satisfactorily. However, if the number of missing values is substantial a serious bias may be introduced. A second approach is to use imputation, i.e., to insert fictitious values where observed values are missing. Since imputed values have a different status from observed values, standard analysis must be modified in order to account for this difference. A third approach is to apply some kind of weighting technique. Weights are assigned to available observations so that the weighted observations are adjusted for missing observations. A fourth and final approach is to define a model for the mechanism that causes missing observations. Inference based on such a model can account for the missing values, but the validity of the model should, of course, be checked.

Vital for proper treatment of missing data is whether or not there is a relation between the missing data mechanism and the variable to be investigated (Y). It is very important to have auxiliary variables (X) which are not affected. In the first part of the book three cases are distinguished: (1) the missing data mechanism depends on Y, and possibly X as well, (2) the mechanism depends on X but not on Y, and (3) the mechanism is independent of X and Y. Case (3) is denoted by "missing completely at random" (MCAR), and case (2) by "missing at random (MAR) within subclasses of X." This case corresponds to what is called an "ignorable" missing data mechanism.

Chapter 2 reviews methods to take care of missing values in the dependent variable in ANOVA. A straightforward method would be to discard cases with missing data and carry out the analysis on the remaining, complete cases. This approach has the disadvantage that balanced designs become unbalanced, and this increases the computational effort considerably. A better approach is to estimate the missing values, and to carry out the analysis on all cases. Various techniques are discussed.

Chapter 3 concentrates on multivariate analysis. Here, a complete case analysis works well only under MAR, otherwise estimators may have a severe bias. Furthermore, the number of discarded cases may be so substantial that hardly any are left for analysis. Available case methods use all available cases for a particular estimator. A consequence is that different estimators may be based on different sample sizes and different cases, which complicates the comparison of results. Furthermore, estimates of correlations may be outside [–1, 1], and covariance matrices may not be positive definite. Another approach is to fill in missing values (imputation). Mean imputation and regression imputation may produce unbiased estimators, but estimated variances are often not correct. The authors do not recommend these methods. Their performance is unreliable, and the required adjustments are too ad-hoc to yield satisfactory results.

Chapter 4 discusses methods to deal with nonresponse in survey sampling. Two different ways of nonresponse modeling are used interchangeably, which is a little confusing. Sometimes the "fixed response approach" is used in which the finite population is divided in a response and a nonresponse stratum, whereas at other points each element in the population is assumed to have a certain, unknown probability of response (the "random response approach"). Two approaches to nonresponse adjustment are considered: weighting cell adjustment and imputation. A special but important case of weighting, raking, receives little attention. Recent results with respect to the variance of raking estimators are not mentioned. Several types of imputation are discussed (mean, random, regression, hot deck, cold deck). Many of the imputation methods have problems with variance estimation. Therefore, some attention is paid to variance estimation based on ultimate clusters. One practical problem with imputed data sets is not discussed, namely, that other users of the data set may believe that they have a complete data set. Analysis of such data sets, particularly on certain selected subsamples or variables, may produce erroneous results. Another approach to tackling nonresponse is not mentioned here: resampling of nonrespondents. Resampling produces indications of whether estimators are biased, and these results can be used to correct for bias.

The second part of the book presents a systematic approach to the analysis of data with missing values. Inference is based on maximum likelihood (ML) techniques derived from statistical models for the data and the mechanism causing data to be missing. Chapter 5 starts with a summary of the maximum likelihood theory for the complete case. It is shown that for large samples and under assumed asymptotic normality, this approach also produces valid results for the incomplete case under MAR. Another approach is to treat the missing data as parameters which can be estimated using ML methods. In practice, this approach is only useful if the fraction of missing values tends to zero.

Chapter 6 discusses methods based on factoring the likelihood so that each factor corresponds to the likelihood for a complete case or for an easier missing data problem. An interesting result is that for the case of bivariate normality, this method is the same as regression estimation, a common estimation technique in survey sampling.

Since in general, missing data patterns are not simple, ML estimation is not always possible. Chapter 7 considers iterative methods to obtain estimates. Special attention is paid to the expectation maximization (EM) algorithm. The idea is to alternatively estimate the missing data and the parameters of the model until convergence. A practical problem in applying this algorithm is the convergence speed, which is proportional to the fraction of missing data.

In Chapters 8, 9, and 10, ML and EM techniques are applied to various problems involving incomplete data. Chapter 9 concentrates on continuous variables, under the assumption of multivariate normality. Applications considered are factor analysis, variance components, linear regression, and time series. Chapter 9 deals with categorical variables, under multinomial and log-linear models. Chapter 10 discusses mixed normal and categorical data. In all cases, the assumption of ignorability proves vital.

The final chapter is devoted to a model-based approach to nonresponse in sample surveys. Special attention is paid to multiple imputation.

The book is not restricted to the missing data problem in survey sampling, and there-

fore, not all theories and applications within this field are discussed. For those concerned with sample surveys only, Madow, Nisselson, Olkin, and Rubin (1983) is an important additional source. Still, the book gives a very good overview of the missing data problem in general, and the state of the art in dealing with missing data. It would be valuable, however, to see the techniques implemented and tried on real data. It is not a simple book. A good background in many aspects of statistics is necessary to understand all subjects treated. Many estimation procedures are presented and discussed, but practical statisticians may need more concrete algorithms or references to existing computer programs. The book should be studied in the statistical methods department in every statistical agency.

References

Madow, W.G., Nisselson, H., Olkin, I., and Rubin, D.B., Eds. (1983): Incomplete Data in Sample Surveys (Vols. 1–3). Academic Press, New York.

Jelke G. Bethlehem
Netherlands Central Bureau of Statistics
Voorburg
The Netherlands

McLachlan, G.J. and Basford, K.E., Mixture Models: Inference and Applications to Clustering. Marcel Dekker, Inc., New York, 1988, ISBN 0-8247-7691-7, xi + 253 pp., $ 69.75 (U.S. and Canada), $ 83.50 (all other countries).

In the last twenty years the quantity of literature on cluster analysis has increased dramatically and a number of books have appeared on the subject including those of Hartigan (1975), Everitt (1980), and Gordon (1981). Each of these has given brief descriptions of an approach to clustering based on finite mixture distributions, particularly mixtures of normal distributions. The text by McLachlan and Basford provides a more comprehensive account of this approach to clustering.

After a brief introduction to the history of mixture models, the authors quickly start their discussion of estimation via the expectation-maximization (EM) algorithm, identifiability, and tests for number of components. Chapter 1 also contains an account of the properties of maximum likelihood estimators including the computation of the associated information matrix. Later chapters cover normal mixtures in some detail. Other types of mixtures such as latent class distributions are covered in somewhat less detail, as are the estimation of mixing proportions, the partitioning of treatment means in ANOVA, and the maximum likelihood approach to the clustering of three way data. Much of the material is technically demanding; however, the many interesting numerical examples provided are helpful in clarifying the more difficult points. A topic not covered but one which would be of practical importance is mixture models for data containing both continuous and categorical variables. The text contains a valuable bibliography with approximately 450 references, most of which are relatively recent, and a number of FORTRAN listings which may be of interest to researchers.

On the whole, the book is well written and well produced and contains much that will be of interest to applied and research statisticians and to those in other disciplines involved in the application of clustering techniques. The text's use for students is limited both by the level of the material and by the lack of suitable exercises.

There are two other texts covering mixture distributions which might be considered competitors – Everitt and Hand (1981), and Titterington, Smith, and Makov (1985). McLachlan and Basford's book is, however, different from both, since it concentrates on the use of mixture distributions as models for the classification process. The book would be a useful addition to the library of those statisticians interested in the theory or application of clustering.

References

Everitt, B.S. (1980): Cluster Analysis. Second edition, Gower Press, London.

Everitt, B.S. and Hand, D.J. (1981): Finite Mixture Distributions. Chapman and Hall, London.

Gordon, A.D. (1981): Classification. Chapman and Hall, London.
Hartigan, J.A. (1975): Clustering Algorithms. Wiley, New York.
Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985): Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Brian S. Everitt
University of London
London
U.K.

Searle, S.R., Linear Models for Unbalanced Data. John Wiley & Sons, Inc., New York, 1987, ISBN 0-471-84096-3, xxiv + 536 pp., $ 49.95.

Searle's popular and successful (1971) text has introduced a generation of students to the linear models which underlie factorial experiments. It advocates use of the overparameterized singular linear model, which has the virtue of containing easily interpreted individual parameters for all main effects and interactions, in exchange for considering such concepts as generalized inverses of matrices and estimable vs. non-estimable parameters. Other distinguishing characteristics of the (1971) book are its devotion to the analysis of unbalanced data (unequal numbers of observations in the sub-most cells), and an unusually clear and meticulous (though at times verbose) writing style. The estimation and testing procedures in the widely-used SAS-GLM computing procedures owe a heavy debt to Searle's approach in the (1971) text.

Not surprisingly, the (1971) book has become dated by subsequent developments. Despite its many strong points, its continued use as a text has become difficult to justify. Chief among these newer findings has been the promotion since the mid 1970s by R. R. Hocking and others of the superiority in most contexts of the full-rank cell means approach to linear modeling.

With the publication of Hocking's (1985) linear models text, there was at last an attractive up-to-date alternative to Searle (1971). However, the former is written at a somewhat higher level of mathematical sophistication and abstraction. Searle's new text fills the gap. It is current, and as expected, contains his lucid writing throughout. And Searle now also generally favors the cell means model.

The present text is, in a sense, two books in one. The first five chapters cover modeling and analysis of one-way and two-way fixed effects layouts having unequal (and possibly zero) cell sizes, using elementary algebra and no matrix notation. Analyses with covariables in the one-way set-up are studied in Chapter 6, and Chapter 7 discusses results in matrix algebra and quadratic forms used in the remainder of the book. To this point the presentation is less demanding and sophisticated than that in Searle (1971), but beginning with Chapter 8, which introduces the General Linear Model, the level of the material roughly parallels that of Searle's earlier text. Chapter 9 considers the analysis of the two-way layout using various overparameterized models, while Chapter 10 examines the cell means model for general layouts as well as various special cases, including the possibility of empty cells and the nonexistence of selected interaction terms. General models containing covariables are explored in Chapter 11. Chapter 12, containing 27 pages, relates the material presented to the output from several computing packages, including BMDP, GENSTAT, SAS and SPSS-X. The final chapter is a 33-page survey of mixed models.

Most chapters conclude with a good selection of exercises. Many of these require the completion of algebraic details from the text, while others involve playing through the algebra with artificial integer data selected to minimize computational difficulties. A solutions manual for the exercises is not yet available.

At times, some unfamiliar but sensible matrix notation is employed. As an example, $\{_r u_i\}$ denotes a row vector having ith entry u_i, while $\{_d A_i\}$ represents a block diagonal matrix whose ith block is the matrix A_i.

For purposes of interval estimation and tests of hypotheses, only normally distributed errors are considered. Except for the brief final chapter, the presentation is limited to fixed effects, and Searle does not discuss Generalized Linear Models in the sense of McCullagh and Nelder (1983).

This text has a number of distinctive fea-

tures. Its devotion to the analysis of unbalanced data is appropriate in view of the prevalence of such data in statistical practice. The material is current with a number of newer and recent results, many of them Searle's own contributions. The book is very well written, though some readers already familiar with the material may complain that, especially in the earlier chapters, it is, at times, exactingly clear and painfully redundant. However, my experience from using the (1971) text is that students appreciate and learn from such presentations.

Searle is especially disturbed that statistical computing packages supply a great many statistical tests without precisely displaying the hypotheses being tested. Throughout this book, for all conceivable hypotheses of interest with the models studied, he associates test statistics with hypotheses. It is shown that many frequently available tests involve the cell sample sizes, and it is argued that such data-dependent hypotheses are usually inappropriate.

Section 5.6 discusses how to sensibly select interaction contrasts of likely interest when there are empty cells in two-way layouts. In Chapter 6, the presentation of models with covariables in the one-way case includes tests of the equality of the mean response over all groups, where the groups may have differing mean values of the covariable. The authoritative discussion in Chapter 10 of analyses with more than two factors where all factors are crossed, (as opposed to some nested), goes well beyond both Searle (1971) and Hocking (1985). For example, a theorem (p.407) indicates a necessary and sufficient condition for estimability of a general linear function of cell means in a model with linear restrictions (e.g., those deleting selected interactions) and some empty cells. This is a matter of some delicacy because the incidence of the empty cells affects which interaction contrasts can be estimated. Chapter 11 provides, in great detail, procedures for testing every imaginable hypothesis in models with covariables, and likewise, precisely what is being tested by every conceivable test statistic in commercial software appropriate for analyzing such models. (Hocking (1985) contains relatively little material dealing with covariables.) Chapter 13 is a good current summary of procedures for point estimation in mixed models without delving as deeply into this area as Hocking's book.

While the text material is interesting and appropriate, there are a number of other things that could have been added to this book that would have increased both its quality and length.

Most importantly, the presentation lacks real (or at least realistic) data sets and case studies–the data are all artificial. As a result, the reader cannot completely learn from this text how to proceed to select an appropriate model and how to undertake the fine detail work involved in exploring the sources of significant main effects and interactions. Test results are described as either significant or non-significant, with no reference to α level or P-value; the weight of evidence of test results is not considered. Nor are problems of simultaneous inference dealt with adequately. There is minimal discussion of analysis of simple effects and interpretation of interactions; it is difficult to communicate the essentials without real examples. There is an over-emphasis on hypothesis testing and point estimation at the expense of interval or regional estimation. Instructors who use this book as a course text should address these shortcomings.

There is almost no commentary on the desirability to check for the aptness of assumptions. Many users of these models (and software developers) do not appreciate that the residuals from least squares fits to the types of models considered in this book can and often should be subjected to the same kinds of residual analyses now customarily undertaken in multiple regression contexts: plotting and other techniques to check for outliers, normality, independence, homoscedasticity, etc. Corresponding remedies (transformations, alternative robust procedures, nonparametric or Bayesian approaches) are likewise not considered here.

As a result of the minimal attention given to mixed models, important topics usually considered under the heading "linear models", such as repeated measures and split plots, are not found in this presentation.

Though Searle's command of the algebra is impressive, algebraic details are overemphasized. Many of those contained in the first six chapters are uninteresting, and mostly unnecessary for those who will proceed on to the

general matrix formulations in the later chapters.

On the other hand, there is practically no mention of the underlying geometry of least squares and linear models, much of which requires minimal mathematical background and is helpful for a thorough appreciation of the statistical analyses.

A curious exception to these impressions is Searle's examination, for additive models, of whether the pattern of empty cells creates a disconnected layout. He presents the geometric approach of Weeks and Williams (1964), which is fine for small two-way layouts though unhelpful in larger situations, but forgoes the opportunity to present the more generally useful algebraic condition involving the rank of the coefficient matrices of reduced sets of normal equations. Hocking's (1985) approach to disconnectedness, which uses a certain canonical form of the model, is a better approach than Searle's for addressing this issue.

The notion of the basis of a vector space is rather elementary. Thus it is disappointing that Searle talks about there being at most rank (X) linearly independent estimable functions of the parameter vector without referring to the notion of a *basis set of estimable functions*. On page 354, he gives an example where knowledge of a basis set of interaction contrasts can bypass the need to invert a certain matrix. In general, if all else fails, the Gram-Schmidt orthogonalization can be used to construct such a basis set. Furthermore, Section 12.3, which discusses how SAS-GLM generates basis sets of estimable functions, could have been shortened with the use of basis terminology.

From my reading, I suspect that this book is as laden with typographical errors as were the early printings of Searle (1971); these will undoubtedly be corrected in future printings. Not all such errors are trivial–there are two different errors in what is arguably the most important formula in the book, (146) on p. 291, which gives the F statistic for a general linear hypothesis.

I recommend consideration of this text for a one or two semester course in linear models for factorial experiments. If you have successfully used Searle (1971) for this purpose, here is its replacement. Many errors and misconceptions in the earlier book have been cleaned up, and the new version is current, correct, and well-written. However, as noted above, the text material needs to be supplemented in certain areas, especially with examples that thoroughly discuss the analysis of data from some real experiments.

The book is also an important reference for anyone needing to analyze designed experiments with continuous responses and unequal numbers of observations in the sub-most cells where these models may additionally include covariables. It is the best available source for learning about the correspondence between the test statistics and hypotheses inherent in the generation of computer output and data analysis for such experimental situations.

Chapters 10 and 11 contain much new and recent material that suggests ideas for research in linear model theory. However, for students intending to do doctoral-level research in the areas of linear models dicussed in this book, assuming that only a single textbook is to be adopted, Hocking (1985) would be a better choice.

References

Hocking, R.R. (1985): The Analysis of Linear Models. Brooks/Cole Publishing Company, Monterey, CA.

McCullagh, P. and Nelder, J.A. (1983): Generalized Linear Models. Chapman and Hall, London.

Searle, S.R. (1971): Linear Models. John Wiley and Sons, New York.

Weeks, D.L. and Williams, D.R. (1964): A Note on the Determination of Connectedness in an N-Way Cross Classification. Technometrics, 6, pp. 319–324.

Burt S. Holland
Temple University
Philadelphia, PA
U.S.A.

Printed September 1988.

Journal of
OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

Mailing address: Journal of Official Statistics, Statistics Sweden, S – 115 81 Stockholm, Sweden

Telephone: +46 8 782 94 53 (Chief Editor Lars Lyberg)
+46 8 783 40 00 (Statistics Sweden)

Telex: 15261 swestat s

Four issues per year
Subscription rates:
Volume 4, 1988. Price USD 30 worldwide (SEK 250 in Sweden). Current single copy price, USD 10 (SEK 70). For members of the American Statistical Association (ASA), the International Association of Survey Statisticians (IASS) and the International Association for Official Statistics (IAOS) the journal is available at USD 24 (SEK 200).

Subscription address:
Journal of Official Statistics
SCB-Distribution, S-701 89 Örebro, Sweden

Change of address:
Send new address with old address label to the Subscription address.

Aims and Scope

The *Journal of Official Statistics* is published by Statistics Sweden, the national statistical office of Sweden. The journal publishes articles on methodology and policy related to statistics produced by national offices and other statistical organizations in both industrialized and in developing countries. The journal focuses on the problems of national statistics production.

We encourage articles on the following topics.

1. Methodology useful to government statisticians

This is to be understood in a wide sense, including methodology for collecting, processing, analyzing, presenting and distributing statistical data. For example, sampling design, estimation, and analytical uses of data fall under this heading, as well as questionnaire design, quality control, data base management, confidentiality, evaluation, and identification of statistical needs. Articles may present new methodology, interesting applications of existing methodology, a comparative study of different methods, or an authoritative exposition of existing methods in a certain field.

2. Policy issues in national statistics production

This includes the relations of the statistical office to the general public, users of statistics, and other authorities. Articles will discuss, for example, statistical programs, dissemination of statistics, presentation of quality, training of statisticians, ethical issues, and the role of statistics in society.

Articles should be as concise as possible without loss of clarity. Applications should be emphasized.

Under the heading *Miscellanea* the journal publishes informative essays dealing with topics considered to be of general interest.

Letters to the Editor are confined to discussions of papers which have appeared in the *Journal of Official Statistics* and of important issues facing the statistical community.

The journal also contains *Book Reviews*.

An *Index* appears in the last issue of each volume.

Information for Authors

See inside back cover.

Submission of Manuscripts

Papers submitted for publication should be in final form and sent to the Chief Editor. Books to be reviewed should be sent to the Book Review Editor. It is assumed that a submitted manuscript has not been previously published and is not under consideration by any other journal. Manuscripts should be submitted in five copies. All manuscripts are refereed. The language of the journal is English.

Authors are fully responsible for all information included in their manuscripts, as well as for their written English. The Editor may assist the author in having the language revised.

Authors receive 50 and book reviewers 20 offprints free of charge.

Preparation of Manuscripts

Publication Format. The journal page contains 13.5 cm by 20.2 cm of printed text. The author may estimate the published length of his/her paper by calculating approximately two typescript pages per journal page.

Manuscripts. Manuscripts should be typed on one side of the paper with good margins on all sides (at least 4 cm on the left) and double spaced throughout.

Title. The title should be brief and specific.

Abstract and Key Words. Each article should be accompanied by a short abstract (as brief as possible, 150 words maximum). The key words should directly follow the abstract.

Headings. These should be numbered. Subheadings and sub-sub-headings may be used.

Tables and Graphics. Special care should be taken in preparing drawings for the tables and graphics. Except for a reduction in size they will appear in the final print in exactly the same form as submitted by the author. Tables and graphics should be planned so that they can be reduced to 13.5 cm width. Lettering should take account of the reduction involved.

Each table and graphic should preferably be prepared on a separate sheet. Appropriate placement in the text should be indicated.

Formulas. Since the column width is 6.5 cm only, authors are encouraged to write their formulas so that they conform to this column format, as in the following example:

$$\bar{y}_d^{*\prime} = \Sigma\left(\frac{1}{N}\frac{Y_{hr}}{N_{hr}}\hat{N}_h + \frac{1}{N}\frac{N_h}{N_{hr}}\hat{Y}_{hr}\right.$$

$$\left. -\frac{1}{N}\frac{N_h Y_{hr}}{N_{hr}^2}\hat{N}_{hr}\right). \qquad (2.1)$$

Equation numbers should be included only when equations are referred to; the numbers should be placed on the right. See above.

The author may list on a separate sheet entitled "Special instructions to the printer," any information that adds clarity. Symbols are printed in *italics*. Underlining indicates italics, if not otherwise specified. The journal does not use bold print for Greek letters.

It is important to distinguish carefully between
- Capital and small letters
- Certain Greek letters and similar Roman ones
- Subscripts, superscripts and "ordinary" symbols.

If a letter or symbol might be misinterpreted it should be explained in the margin at its first appearance in the manuscript.

Reference Citations. The JOS style in citing a reference is to use the author's surname and the year of the publication as in the following examples:

Dalenius (1974)
Swensson (1974, 1977)
Bourke (1974a)
Boruch and Cecil (1979, p. 154)
Linebarger et al. (1976)

To distinguish between publications by the same author in the same year use a, b, c, etc.

References in the text should be indicated in the following ways:

1. When a reference is directly cited:
   as discussed by Dodge and Romig (1944)
2. When a reference is cited as an example:
   as discussed previously (Dodge and Romig (1944))
3. When a reference is made to a particular page, section or formula in a work:
   We rely on the algorithm of Dasgupta (1965, pp. 115–120).
   The distribution is known to be normal (Smith and Smith (1958, Ch. 5))

The Reference List. References should be arranged alphabetically and for the same author chronologically.

1. Author's name and year of publication.
2. Title.
3. Details of publication. Complete name of journal. Publisher. Publication site. Page(s) referred to.

References should be given in the standard form of the following examples:

Lininger, C. and Warwick, D. (1975): The Sample Survey, Theory and Practice. McGraw-Hill, New York.

Pomeroy, W.B. (1963): The Reluctant Respondent. Public Opinion Quarterly, 27, pp. 287–293.

Platek, R., Singh, M.P., and Tremblay, V. (1978): Adjustments for Non-response in Surveys. In Survey Sampling and Measurement, edited by N.K. Namboodiri, Academic Press, New York.

National Center for Health Statistics (NCHS) (1965): Health Interview Responses Compared with Medical Records. Vital and Health Statistics, P.H.S. Publication No. 1000 – Series 2 – No. 8. Government Printing Office, Washington, D.C.

Printed by Statistics Sweden, SCB, Frösö, Örebro, Sweden