Journal of Official Statistics, Vol. 13, No. 2, 1997, pp. 103–112

Evaluation of a Reconstruction of the Adjusted 1990 Census for Florida

Michael M. Meyer[1] *and Joseph B. Kadane*[1]

Meyer and Kadane (1992) report a method for reconstructing the adjusted population (by age, race, and sex) for the half of the census blocks in Florida not made available to them. This article studies the full adjusted data set, which is now available, to examine how well the original reconstruction was done. This is a rare opportunity to learn the exact value of quantities estimated.

The results show that the largest difference between the Meyer and Kadane (1992) approximation and the adjusted counts at the Congressional district level was 79 persons for one district. Thus, the approximation could have been used instead of the unavailable adjusted census, had the redistricting decision-makers so chosen.

Key words: Disclosure avoidance; exploratory data analysis; confidentiality; geographic boundary effects.

1. Background

The question of whether to adjust the U.S. Census because of undercounting has been a legal, political, and scientific issue for more than a decade (see Breiman 1994; Freedman and Wachter 1994; Belin and Rolph 1994; and the references cited there). In 1987, the Commerce Department in the Reagan administration announced a decision not to adjust the 1990 census. This subsequently led to the filing of the lawsuit of City of New York et al. vs U.S. Department of Commerce et al., challenging that decision on the basis that it violated the equal protection clause of the constitution with respect to voting rights.

In a partial settlement of this case, the administration agreed to reopen the decision, and to collect and analyze the Post Enumeration Survey data required if adjusted data were to be used. In July, 1991, the Bush administration Commerce Department announced a decision not to adjust the 1990 census on the basis of the 1990 Post Enumeration Survey. In addition, the Commerce Department declined to release the adjusted data at the census block level, thus preventing states from using adjusted data to establish boundaries for congressional and state legislative seats. In the fall of 1991, the Subcommittee on Census and Population of the U.S. House of Representatives Committee on Post Office and Civil Service, under the

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Acknowledgments: The authors thank Michael Daniels, TerriAnn Lowenthal, and Robert Rifkind for helpful conversations. The research was partially supported by NSF Grant DMS-9303557. Figure 1 is reproduced from Preimesberger and Tarr (1993) by permission of Congressional Quarterly, Inc., Washington D.C. This work would not have been possible without the computing facilities available in the Department of Statistics at Carnegie Mellon. We also thank the Associate Editor for very helpful expositional comments.

leadership of Congressman Thomas Sawyer, subpoenaed the block-level adjusted counts; in a subsequent compromise, the Secretary of Commerce agreed to release the adjusted counts for every other census block in the nation and for all aggregations having an adjusted population of at least 1,000.

At this point, we were approached by leaders in the House of Representatives of the State of Florida, who were interested in using adjusted census data for redistricting in Florida if we could "fill in" adjusted numbers for the census blocks withheld by the Commerce Department under its agreement with Congressman Sawyer. Meyer and Kadane (1992) (abbreviated hereafter as MK) records the methods we used to do this, and gave a preliminary evaluation using city adjusted populations, which we did not use in our procedure. Because the Florida Senate did not want to use adjusted census data, our study was not actually used in redistricting Florida. We subsequently learned (Hogan, 1993, p.1054) that there were errors in the Census adjustment procedure. MK aimed to reconstruct the official (even if incorrect) adjusted census and used only minimal meta-knowledge about the adjustment procedure. Hence we have based our comparisons on the published adjusted census.

As part of a 1994 U.S. District Court ruling in the City of New York, et al. vs U.S. Department of Commerce, et al. lawsuit, which challenged the Commerce Department's original ruling on adjustment on constitutional grounds, tapes containing the adjusted census block data for the entire U.S. were released into the public domain. The purpose of this article is to examine how well we did in reconstructing the population of the census blocks for which data were not available under the agreement between Congressman Sawyer and the Department of Commerce. What might have been the consequences had our results been used to redistrict Florida? Section 2 gives an overview of the methods of MK, Section 3 gives results for reapportionment, particularly at the Congressional level. Section 4 decomposes the errors made according to the steps outlined in Section 2. Section 5 reports our conclusions.

2. Methods

The main way that the U.S. Census Bureau organizes its population files is hierarchical. Data are recorded at various levels of aggregation and these aggregation levels are given names like 750, 740, and so on. The most detailed information available is the block level, which is coded as a level 750 area. Each 750 level record is an element of exactly one Block Group, denoted by a 740 level record, which in turn is an element of exactly one Census Tract, a 730 level record, and so on through levels 720, 710, 700, and 060 (an entire county). Additionally, there are certain other summary levels (or cross-aggregations) of the census blocks but these are not part of this hierarchy: 140 levels are also (confusingly) called Census Tracts, 160 levels are named places, 170 are cities, and there are special aggregations for Indian Reservations. All summary levels are built up from basic level 750 (block) units.

MK had available the unadjusted census counts at all these levels, together with the Sawyer tape, containing adjusted counts for every other 750 level area, and adjusted counts for all levels in which the adjusted counts were at least 1,000. For each such geographical area, both the unadjusted and adjusted tapes report the total population and 23 additional numbers, dividing the population by age, race, and ethnic origin, in overlapping

categories. We reformulated these 24 numbers into 20 independent and exhaustive counts, which we call data streams, as detailed in MK, Section 4.

There were two fundamental methods we used to reconstruct the missing block records: exact and approximate. MK used exact methods where possible, and otherwise approximated. The first kind of exact reconstruction used is that when an unadjusted population is zero, then the adjusted number must also be zero. There are many unadjusted zeros, so this trivial method was quite useful. The other exact methods make use of the relationships between areas inherent in the files. These relationships are:

- If area A contains area B (say A is a 740 level area and B is a 750 level area), and the unadjusted count for B is the same as that for A, then so must be the adjusted counts.
- If area A is the union of several areas B_i, and A or only one of B_i's has missing adjusted data, then the last can be determined, by addition if the missing one is A, by subtraction if it is one of the B_i's.

These methods were used iteratively until no further changes were made both on the main hierarchy (750, 740, etc.), and on the 060 level. We held out the 160/170 data, which were used in a preliminary evaluation.

The second fundamental set of methods we used were approximate methods, which we employed only after getting everything we could from the exact methods. The major ideas in the approximate methods were a ratio adjustment (with some modification for very small numbers) to scale imputed totals to externally known totals, and integerization, so that our reported numbers did not include fractional numbers of people. We used a simple rounding method rather than the controlled rounding technique used by the U.S. Census Bureau as described in Cox and Ernst (1982); Causey, Cox, and Ernst (1985); and Cox (1987).

3. Results for Redistricting

Since the Florida House of Representatives hoped to use our results to redistrict, a most important aspect of our evaluation is how well we did for that purpose. Accepting, as we do, the premise that the adjusted counts are most accurate, was the error induced by the incompleteness of the Sawyer tape greater or less than the error induced by using the unadjusted figures?

Detailed census block data are used every ten years for three purposes: to redistrict the state's Congressional seats, and to redistrict the state's own House and Senate. Thus, from the perspective of use, the important first issue is the size of the errors we made in the populations at each of these levels. This is summarized in Table 1.

It is not hard to understand intuitively that the sum of absolute errors would decrease as the numbers of districts decreases; after all, for the single district consisting of the whole

Table 1. Summary of MK errors in estimating the sizes of districts in Florida

Body	Number of districts	Min error	1st quartile	Median	Mean	3rd quartile	Max	Sum absolute errors
Florida House	120	−23	−5.25	1	0	6	19	854
Florida Senate	40	−42	−11	−1	0	8	45	574
Florida Congressional Representatives	23	−66	−9.5	−4	0	9.5	79	422

Table 2. Fit of crude theory to MK errors

Body	Size of body, k	Total absolute error, E	$E/\sqrt{k}$
Florida House	120	854	88
Florida Senate	40	574	91
Florida Congressional Representatives	23	422	78

state of Florida, the error is necessarily zero. However, it would be nice to have some idea, even if a rough one, of how these errors might decrease as the number of districts decreases.

Imagine a state with a uniformly distributed population P, divided into k districts, each with population P/k. Suppose these districts are geographical squares of side s, so $s^2 \propto P/k$. Suppose also that absolute errors e per district occur in proportion to the length of the boundary, so $s \propto e$. Total absolute error $E \equiv ke \propto ks \propto k\sqrt{P/k} \propto \sqrt{k}$, so $E/\sqrt{k}$ should be constant. How well does this work for our three apportionments in Florida? The answer is given in Table 2. We judge from Table 2 that the theory fits reasonably well, given the crudeness of the assumptions.

The most carefully regulated redistricting is Congressional reapportionment, which is heavily influenced by the case of Karcher vs Daggett, 462 U.S. 725 (1983). In that case the U.S. Supreme Court ruled that New Jersey's 1980 redistricting of Congressional seats was unconstitutional because of inequality in the census populations of the districts although these differed by less than one percent.

Table 3 gives three population figures for each of the 23 Congressional seats Florida was entitled to after the 1990 census. This table, and similar ones for the Florida House and Senate are available from StatLib at `http://www.stat.cmu.edu/general/mk-florida`. The fifth column gives the unadjusted census populations, which are within one person of each other, thus complying with the Karcher decision. The second column gives the adjusted populations of these districts, which range from 583,506 in District 21 to 566,964 in District 9, a difference of 16,542, or about 2.8%. Hence failure to use the adjusted census data may make these districts vulnerable to legal attack citing Karcher. Column 3 in Table 3 gives the MK approximation to Column 2, and Column 4 records the error in the approximation. It is obvious that the errors made in our approximation are orders of magnitude less than the errors that give rise to discrepancies between the adjusted and unadjusted numbers.

Districts 3 and 4, which showed the largest absolute errors in the MK approximation, are adjacent to each other. This finding along with comparable-sized errors of opposite sign in other pairings of geographically adjacent districts (19 and 20, 5 and 6, 7 and 8, 21 and 22, and 23 and 18) suggests that perhaps a single large error on the boundary may have caused the errors. To study this hypothesis, we study our errors in Districts 3 and 4 in some detail. Table 4 summarizes the errors made at the Block (750) level in those districts.

Evidently, the error in both districts is the consequence of the balancing of many errors, most of them very small, rather than a single large error. Consider, for example, the largest block-level error in Table 4, the -48 in District 3. If this came from an area in the middle of a compact district, say in a county entirely included in District 3, then there would have

Table 3. Results of redistricting Florida for the 102nd Congress

Dist	Adjusted	MK approx	Error	Unadjusted	Adj−unadjusted
1	579,970	579,967	3	562,518	17,452
2	580,173	580,181	−8	562,519	17,654
3	582,909	582,975	−66	562,519	20,390
4	576,784	576,705	79	562,518	14,266
5	571,545	571,573	−28	562,518	9,027
6	575,058	575,038	20	562,518	12,540
7	579,492	579,507	−15	562,518	16,974
8	581,455	581,436	19	562,518	18,937
9	566,964	566,956	8	562,518	4,446
10	569,395	569,391	4	562,518	6,877
11	577,449	577,456	−7	562,519	14,930
12	576,381	576,385	−4	562,519	13,862
13	574,475	574,470	5	562,518	11,957
14	576,284	576,291	−7	562,518	13,766
15	577,698	577,702	−4	562,519	15,179
16	577,369	577,360	9	562,519	14,850
17	580,571	580,580	−9	562,519	18,052
18	580,561	580,538	23	562,519	18,042
19	576,453	576,422	31	562,519	13,934
20	577,975	578,007	−32	562,518	15,457
21	583,506	583,496	10	562,519	20,987
22	573,547	573,557	−10	562,519	11,028
23	581,694	581,715	−21	562,519	19,175

to be positive errors canceling the −48, and the net error contributed to MK's estimate of the adjusted District 3 population would be zero. In fact, however, that level 750 block is in a level 740 aggregation shared between Districts 3 and 4. The total error in that 740 aggregation is 25, and there is no error in the District 4 part of the level 740 aggregation. The 730 area enclosing the level 740 area that is split between districts has zero error, and contains only one 750 level geographic unit in District 4, which also had a zero error. Hence the net contribution to the Congressional District 3 error from the 48 error is zero: the rest of the level 730 area that contains it balances it out. A similar analysis shows that the −21 error also did not contribute to any aggregate error at the Congressional district level. Thus we are forced to conclude that there is not a single large error between Districts 3 and 4 producing the net error numbers found in Table 3.

Further analysis of census units shared between Districts 3 and 4 supports the impression that the net errors in Table 3 are driven by an accumulation of many small errors. District 3 serves as an interesting case to study because it is far from compact in shape, (see Figure 1); according to one description it "meanders about 250 miles through 14 counties" (Preimesberger and Tarr 1993, p. 170). This peculiar shape apparently arose out of an attempt to comply with the extension of the Voting Rights Act passed in 1982, in which the U.S. Congress outlawed any redistricting practice discriminating about certain minority groups whether intended or not. The current law is in a state of flux on just what effort is required, and permitted, to provide fair representation for minorities (see Preimesberger and Tarr 1993, p. 17, Miller vs Johnson (115 S.Ct. 2475, 1995) and Bush vs Vera (116 S.Ct. 1941, 1996)); for present purposes in analyzing errors from the MK

Table 4. MK errors in Congressional Districts 3 and 4 at the 750 level

Errors	District 3	District 4
Total	−66	79
large:	12,16,17	14
10		1
9		6
8		3
7	2	3
6	6	4
5	11	7
4	17	16
3	44	20
2	177	92
1	932	606
0	12,854	10,528
−1	906	629
−2	95	62
−3	42	9
−4	13	6
−5	9	8
−6	6	2
−7	5	3
−8	4	2
−9	1	1
−10	4	2
small	−11,−12,−13(2),−14(2), −16,−19,−22(2), −27,−48	−11(3),−12,−13 −18,−19

approximation, an important consequence of these rules and of the unusual shape of District 3 is that nearly all of it is close to a boundary with some other district.

To examine more closely the boundary between Districts 3 and 4, Table 5 assembles, for each level of geography, the areas shared between those districts (and no others). Again, no single large error appears. We also looked at various summary levels which have members in three districts, but the errors were so small as not to contribute much to the overall picture given in Table 5.

From these analyses, we conclude that the apparent pattern in Table 3 of pairs of districts with errors of roughly equal magnitude and opposite sign masks a more complex reality. Each net error appears to be the result of many small errors, most of which cancel, rather than a few large errors. In a sense this is heartening, since it suggests that our errors are related not to outliers but to individually small additive errors that might be amenable to a better method.

4. Decomposition of Error

As outlined in Section 2, the MK methods consisted of three steps: (1) imposition (iteratively) of exact restrictions, (2) extrapolation, and (3) integerization. This section investigates how the error is apportioned among these.

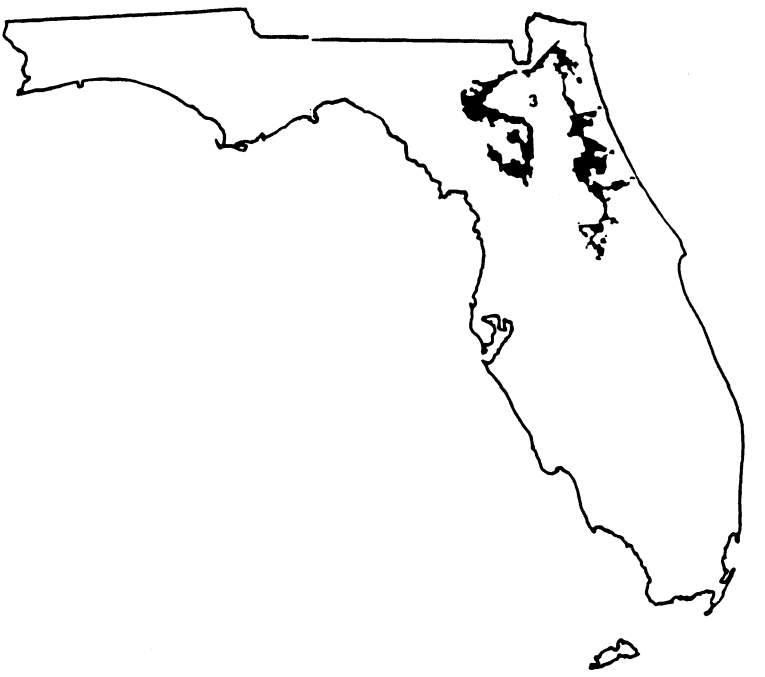

Fig. 1. Florida Congressional District 3

The exact methods rely on deductive inferences stemming from our understanding of the nature of the geographic information. In practice, the "exact" adjustments we deduced agree, exactly, with the adjusted data. Thus, the exact restrictions were performed without error, so the first possible source of error is eliminated.

To examine how our error is properly apportioned between extrapolation and integerization,

Table 5. MK errors on the boundary between Districts 3 and 4 at various levels of the geographical hierarchy

Error	Level				
	740	730	720	710	700
Total	14	1	−15	−15	−15
large	22,15,14,12	36,15,7	0	0	0
	8,7,6	0	0	0	0
5	0	1	0	0	0
4	3	1	0	0	0
3	4	1	0	0	0
2	3	3	0	0	0
1	11	4	2	2	2
0	43	56	69	68	65
−1	10	4	2	1	1
−2	7	4	1	0	0
−3	2	3	1	2	2
−4	2	0	1	1	1
−5	7	2	2	0	0
small	−7, −12, −19, −25	−6, −6, −36	6	6	6

Table 6. Summary of MK errors in estimating the sizes of districts in Florida

Body	Number	Min error	1st quartile	Median	Mean	3rd quartile	Max	Sum absolute errors
Florida House integers	120	−23	−5.25	1	0	6	19	854
real numbers		−9.014	−1.47	0.1388	−0.0001969	1.558	6.233	246.4994
Florida Senate integers	40	−42	−11	−1	0	8	45	574
real numbers		−11.14	−3.822	−0.47	−0.0005908	3.36	13.73	180.1354
Florida Congressional Representatives integers	23	−66	−9.5	−4	0	9.5	79	422
real numbers		−15.96	−2.131	−0.1159	−0.001028	1.633	16.1	100.2998

we imagine that our instructions had been changed so that we could report a real number, rather than an integer, as our estimate of the number of people, or number of people of a particular type, in a geographical area. Even though the correct number must be an integer, reporting as real numbers would reduce the error. To see this, suppose that the extrapolation yields 34.5 persons in a particular level 750 area. To report either 34 or 35 risks a larger error than reporting 34.5. Thus the errors made from a hypothetical problem in which non-integer numbers are accepted allows separation of extrapolation errors from integerization errors.

Table 6 reports the results of this exercise. In comparison to the previously reported integerized results, the results show a dramatic drop in the errors made had real-number reporting been permitted. In terms of the sum of absolute errors, extrapolation accounts for 28.8% of the Florida House error, 31.4% of the Florida Senate error, and 23.7% of the Florida Congressional Representatives error. The largest errors are also smaller with real-number reporting. Consequently it is clear that integerization, not extrapolation, is our biggest source of error.

5. Conclusions

It is unusual in prediction problems to ascertain exactly the numbers one has been trying to estimate. In this instance, a turn in the sequence of events led to our good fortune in obtaining a gold standard for evaluating our original work on Florida's redistricting. Overall, the error in the MK approximation to the adjusted census counts was relatively modest. As shown in Table 1, the errors made for apportionment are quite small, probably smaller than other sources of error in the system.

How might we have improved our work had we been confronted with a similar issue in the future? We considered the possibility that using slightly different extrapolations as a function of minority status (Black, Hispanic, other) might have helped. However, Section 4 shows that most of our error is due to integerization rather than extrapolation. Thus, if greater accuracy were required, it appears that effort should be placed on making our integerization more sophisticated. Our errors in the apportionment context are the sum of many small errors, most having to do with integerization, and roughly explained by our crude theory reported in Table 2.

The intent of the Bush Administration in resisting the Sawyer subpoena for the adjusted census at the block level, and in agreeing to release only half of it, was apparently to make it impossible to use the adjusted census of 1990 for redistricting. Dedicated examination of the remaining clues allowed MK to do quite well in restoring the original numbers. Others, attempting to conceal microdata, perhaps for privacy or confidentiality purposes, might think again about how well they have succeeded. Harry Roberts (1986) in his comment on the ground-breaking Duncan-Lambert article on disclosure limitation (1986) remarks "the tools of analysis that can invade individual privacy can be used to defeat government attempts at censorship and concealment." This article documents our success in overcoming one such attempt.

6. References

Belin, T.R. and Rolph, J.E. (1994). Can We Reach Consensus on Census Adjustment? Statistical Science, 9, 486–504.

Breiman, L. (1994). The 1991 Census Adjustment: Undercount or Bad Data? Statistical Science, 9, 458–475.

Cox, L.H. and Ernst, L.R. (1982). Controlled Rounding. INFOR, 20, (4), 423–432.

Causey, B.D., Cox, L.H., and Ernst, L.R. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, (392), 903–909.

Cox, L.H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. Journal of the American Statistical Association, 82, (398), 520–524.

Duncan, G.T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination. Journal of the American Statistical Association, 81, 10–28 (with Discussion).

Freedman, D. and Wachter, K. (1994). Heterogeneity and Census Adjustment for the Inter-censal Base. Statistical Science, 9, 476–485.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. Journal of the American Statistical Association, 88, (423), 1047–1060.

Meyer, M.M. and Kadane, J.B. (1992). Reconstructing the Adjusted Census for Florida: A Case Study in Data Examination. Journal of Computational and Graphical Statistics, 1, 287–300.

Preimesberger, J. and Tarr, D. (eds.) (1993). Congressional Districts in the 1990's: A Portrait of America Congressional Quarterly, Inc.: Washington, DC.

Roberts, H. (1986). Comment. Journal of the American Statistical Association, 81, 25–27.

Received November 1995
Revised August 1996