SSOAR

Open Access Repository
www.ssoar.info

Problems and opportunities in the use of individual and aggregate level census data

Vinovskis, Maris A.

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Vinovskis, M. A. (1980). Problems and opportunities in the use of individual and aggregate level census data. In J. M. Clubb, & E. K. Scheuch (Eds.), *Historical social research : the use of historical and process-produced data* (pp. 53-70). Stuttgart: Klett-Cotta. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-326415

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Maris A. Vinovskis

Problems and Opportunities in the Use of Individual and Aggregate Level Census Data

Ever since the first federal census was taken in 1790, scholars have used these data to try to understand American demographic and socio-economic development. Most of the nineteenth-century studies of census data were descriptive rather than analytical. During the past two decades, however, historians increasingly have used these nineteenth-century censuses to reconstruct American society from a more quantitative perspective.

Despite the increased usage of census data at both the aggregate and individual levels, very little effort has been made to assess the strengths and weaknesses of this source of information. Most scholars simply have used these data without really considering any of the problems inherent in them[1]. Other historians have altogether neglected this valuable source because they are unaware either of the information found in the censuses or of the analytical techniques available to investigate them[2].

In this essay, I will consider very briefly some of the opportunities and problems in the use of American census data from the nineteenth century. I will first discuss some of the issues raised by the use of aggregate census data and then turn to the problems of using them at the individual level.

[1] This is particularly true of economic historians who tend to use census data without considering the possible biases in such data. For example, the recent studies of wealth inequality in nineteenth-century America have used census information without ascertaining exactly what is being measured by the questions relating to property. Soltow, Lee, Men and Wealth in the United States, 1850–1870, New Haven 1975; Soltow, Lee, Patterns of Wealthholding in Wisconsin Since 1850, Madison/Wisconsin 1971.

[2] Though most social historians now use census data in their analyses, many of them underutilize the available information. Anthony Wallace's recent study of an industrial community uses census information as part of its analysis, but it fails to use those data to explore in more depth the lives of the inhabitants of that community. Wallace, Anthony F. C., Rockdale: The Growth of an American Village in the Early Industrial Revolution, New York 1978.

I. Aggregate Use of Census Data

Scholars from very different perspectives have utilized aggregate census data to investigate our past. Political historians have studied patterns of electoral behavior at both the state and country levels using the federal censuses in conjunction with the national and state election returns. Fertility differentials and trends have been studied by demographic historians while economic historians have analyzed nineteenth-century economic development. Finally, social historians are now beginning to investigate such issues as school attendance, literacy, and urban development.

The first American censuses were gathered primarily for political purposes. As part of the „Great Compromise" of the Convention of 1787, each state received equal representation in the Senate while the House of Representatives was apportioned on the basis of the population of each state. Therefore, it was necessary to provide for some mechanism for counting the population. As a result, a census of the population was instituted for 1790 and held at ten-year intervals thereafter[3].

The first federal census was not very comprehensive. Information was gathered only on six items:

1) The name of the head of the household
2) The number of free white males of 16 years and upwards, including the head of the household
3) The number of free white males under 16 years
4) The number of free white females, including the head of the household
5) The number of all other free persons
6) The number of slaves

It was not until 1830 that regular census schedules were printed and distributed to the census marshalls. Up to that time, each census enumerator simply made their own forms. Combined with the fact that the position of a census marshall was often reserved for rewarding politicans, it is not surprising that the early censuses were not as carefully and accurately carried out as their modern counterparts.

Information for the federal censuses from 1790 to 1840 were collected only at the household level. Therefore, it is impossible to obtain individual level data for any of these years. Beginning with the federal census of 1850, however, the individual became the object of enumeration – a major innovation and improvement in the quality of the data. Furthermore, by that date the questions asked of the population had been greatly expanded to include additional information on such issues as occupation and wealth. Altogether, six separate census schedules were used in 1850 to obtain data on population (both free and slave), agriculture, industry,

[3] For a useful introduction to the development of the federal censuses, see Wright, Carroll D., and Hunt, William C., The History and Growth of the United States Census, Washington D.C. 1900.

mortality, and social statistics (i. e. schools and colleges, libraries, newspapers and magazines, religion, crime, poverty, and wages). Thus, during the course of the nineteenth century, the federal censuses became more detailed and comprehensive both in the type of information they solicited and the manner in which it was recorded.

Much of the census information gathered during the nineteenth century was aggregated at either the state or county level and the results were published by the federal government. These published aggregate data have become the basis of the ecological analyses of census materials. Analyses of these data are greatly facilitated by their availability at the state and county levels in machine readable form through the Inter-University Consortium for Political and Social Research[4].

There are problems, however, with the manner in which the census data were aggregated. The unit of aggregation is usually either the state or the county. But state and county level data often mask significant socio-economic variations within units. Some scholars have used township data for their analyses even though it has meant going back to the original census manuscripts and reaggregating the data or going to some other source such as the state censuses which sometimes are aggregated at the township level. Simply to assume that the proper level of analysis is the state or county because of the lack of data at the township level is a mistake. Some of the most interesting analyses in fields such as educational development need to be examined at the township rather than county level since local variations in the pattern of school attendance or school expenditures often can be quite sizable within counties.

Another problem of using the available aggregate census data is that the units are of such varying size. For example, the population of the largest state in 1850, New York, was 3,048,325 people, while that of the smallest state, Florida, was only 47,203 people. Similarly, whereas the county of New York (New York) had a population of 515,547 people, there were only 79 inhabitants in the county of Clarke (Iowa) in 1850. Anyone doing ecological analysis has to decide whether to use these units as equivalent to each other or to weight them by some factor such as their population size. The correct procedure, of course, depends on the model being tested by our analysis. In most historical work, analysts have treated all units as equivalent and therefore have not weighted them. Generally, this is a conceptually defensible procedure, but sometimes we do encounter serious problems if one of the smaller units has an extreme value.

Let me illustrate this latter point by referring to a study of white fertility ratios at the state level in ante-bellum America. Colin Forster and G. S. L. Tucker analyzed fertility differentials using all the states and territories in the United States between 1800 and 1860. The total number of units in their analysis was always quite small since there are only a limited number of states and territories in this period. Thus, in 1860 there were only thirty-four states and seven territories in the

[4] For a catalogue of the available machine-readable data at the Inter-University Consortium, see Inter-University Consortium for Political and Social Research, Guide to Resources and Services, 1977–1978, Ann Arbor/Michigan 1978. These Guides are updated annually.

United States (not counting the District of Columbia)[5]. As a result, one has to be particularly careful that none of the states or territories have extreme values in any of their variables since this might distort the entire analysis.

In their analysis of 1860, Forster and Tucker used all of the states and territories. Unfortunately, this meant including the Dakota territory which has a white population of only 2,576 in that year. Since the Dakota Territory also had a very low white refined fertility ratio (1,157) and an unusually high white sex ratio (1,710), it seriously skewed their overall results. Rather than have the entire analysis distorted by such an extreme case, they should have either eliminated all of the territories or weighted their states and territories by population size so that their results would be less affected by a new and unsettled area that had such unusual characteristics compared to the other states[6].

The above example also highlights one of the major problems of using aggregate census data at the state level — the small number of cases available. With only thirty-four states in 1860, it is very difficult to use any elaborate regression equation since the degress of freedom lost by each additional independent variable restricts our analysis. Furthermore, any extreme value in any of the cases, whether they be for small or large units, can greatly affect the results since there are so few cases in the analysis. Finally, it is almost impossible to analyze variations within regions of the country using state level data because of the small number of states within any region. Unfortunately, there are examples in the historical literature of correlation analysis being done with such small number of states[7].

In analyzing aggregate census data over time, a further complication arises due to changes in the boundaries of the units. This is particularly true at the county level in the nineteenth century. In many of the newer states, when the population of the counties increased, they were subdivided into smaller units. As a result, direct county to county comparisons over time are difficult for nineteenth-century America.

The use of aggregate census data is also hindered because census units do not always coincide with political units. For example, congressional districts in nineteenth-century America sometimes split counties as the state legislature was more concerned about the characteristics of the voters within each district than about keeping county lines intact. As a result, it is difficult to create files of political and census data at the congressional level.

[5] Forster, C., and Tucker, G. S. L., Economic Opportunity and White American Fertility Ratios, 1800–1860, New Haven 1972.

[6] For a critique of their analysis, see Vinovskis, Maris A., Socio-Economic Determinants of Interstate Fertility Differentials in the United States in 1850 and 1860, in: Journal of Interdisciplinary History, 6 (1976), pp. 375–396; Vinovskis, Maris A., Recent Trends in American Historical Demography: Some Methodological and Conceptual Considerations, in: Annual Review of Sociology, 4 (1978), pp. 603–627.

[7] Some of the analyses at the state level in Yasukichi Yasuba's work are done with only ten or fifteen cases. Yasuba, Yasukichi, Birth Rates of the White Population in the United States, 1800–1860: An Economic Study, Baltimore 1962.

In addition to considering the problems of the comprehensiveness of the census questionnaires as well as the size, number, and comparability of the aggregate census units, we also need to evaluate the quality of the census returns. Most scholars using historical censuses have not paid sufficient attention to the accuracy of the censuses. Usually they simply assume that these data are accurate or that if they are under-enumerated, the degree of under-enumeration is relatively uniform so that comparisons across units are still meaningful.

The quality of census data varies not only by the particular census which is being used, but also by what issue is being considered. Generally, the later censuses are more accurate than the earlier ones since more standardized procedures were introduced for gathering and processing the data. There are some censuses, however, such as that of 1870, which are considered to be inferior in quality to their predecessors.

The data on the general characteristics of the members of a household are considered to be reasonably accurate, though it is clear that some households have been omitted entirely – particularly those which were the most transient. Though scholars are still concerned about the accuracy of these data, most of them are quite willing to accept and use them in their investigations[8]. Other areas of census inquiry are so clearly unreliable that everyone agrees that these data should not be used in further analyses. For example, the data on the insane population are clearly under-enumerated to such a large extent that even contemporary observers such as Edward Jarvis argued that they were worthless. The nineteenth-century federal census marshalls were negligent in collecting the information on insanity and the families they interviewed were often reluctant to admit that a member of their household was insane[9]

Though there is a consensus among scholars on the relative reliability of some categories of census information, there is considerable difference of opinion on other areas such as mortality information. Most demographers and economists continue to use mortality data from the census of 1850 even though other scholars have argued that these data are so badly under-enumerated that any conclusions based on them are highly suspect. Some defenders of the use of these mortality data argue

[8] A few historians have tried to consider the accuracy of the census data. For example, several scholars have investigated the accuracy of age reporting in the federal censuses. Knights, Peter R., Accuracy of Age Reporting in the Manuscript Federal Censues of 1850 and 1860, in: Historical Methods Newsletter, 4, No. 3 (June 1971), pp. 79–83; Hammarberg, Melvyn, The Indiana Voter: The Historical Dynamics of Party Allegiance During the 1870's, Chicago 1977, pp. 210–217.

[9] For a discussion of the under-registration of the insane in the early federal censuses, see Grob, Gerald N., Edward Jarvis and the Federal Census, in: Bulletin of the History of Medicine, 50 (Spring 1976) pp. 4–27; Rosenkrantz, Barbara G., and Vinovskis, Maris A., The Invisible Lunatics: Old Age and Insanity in Mid-Nineteenth-Century Massachusets, in: Spicker, Stuart F., et al. (eds.), Aging and the Elderly: Humanistic Perspectives in Gerontology, Atlantic Highlands/N.J. 1978, pp. 95–125.

that even if these data are under-enumerated, the relative under-enumeration is about the same so that these data are still appropriate for comparative purposes. But there is no reason to believe that mortality data are under-enumerated at the same rate throughout the country or among different segments of the population. In fact, the superintendent of the census, as well as other nineteenth-century scholars, argued that the mortality returns were so badly and unevently under-enumerated that they were of little value even for comparative purposes.

> The federal census of 1850 furnishes the first instance of an attempt to obtain the mortality during one year in all of the States of the Union, and had there been as much care observed in the execution of the law as was taken in framing it, and in the preparation of the necessary blanks, a mass of information must have resulted relating to the sanitary condition of the country, attained as yet in no other part of the world . . . The varying ratios between the States, as drawn from the returns, show not so much in favor of or against the health of either, as they do, in all probability, a more or less perfect report of the marshalls. Thus it is impossible to believe Mississippi a healthier State than Rhode Island, etc.[10]

Thus, one of the major criticisms that can and should be raised against many of the scholars using nineteenth-century aggregate censuses is that not enough attention has been paid to the quality of the data. Much of the recent work in this area is quite sophisticated from a statistical perspective, but very elementary in terms of considering the possible biases in the data. One simply cannot use nineteenth-century census materials as if they were as accurate as those produced by the U. S. Bureau of the Census today.

One other point needs to be mentioned in regard to the accuracy and representativeness of census data. Even if the census data are accurate, it does not mean that they are typical or representative of that decade. Almost none of the studies using census data have adequately tested the possibility that the year in which the census was taken was atypical in terms of the variable being investigated. This is particularly hazardous for historical data since there were often much larger and more frequent fluctuations in demographic and socio-economic variables in the past than today.

Again, let me illustrate this point by using mortality data from the federal census. One of the major life tables for nineteenth-century America is the Jacobson Life Table. It is based on census mortality data for Massachusetts and Maryland in 1850. The data were originally assembled and converted to life expectancies by L. W. Meech and recalculated by Paul Jacobson a hundred years later[11]. The Jacobson Life Table has been accepted by virtually all scholars as the definitive estimate of life expectancy in mid-nineteenth-century America.

[10] U. S. House of Representatives, Mortality Statistics of the Seventh Census of the United States, House Executive Documents No. 38, 33rd Congress, 2nd Session, Washington D. C 1855, p. 8.

[11] Jacobsen, Paul H., An Estimate of the Expectation of Life in the United States in 1850, in: Milbank Memorial Fund Quarterly, 35 (1957), pp. 197–201.

The question we need to ask is whether the period June 1, 1849–June 1, 1850 was a typical year in terms of mortality. Naturally, since the census itself covers only one year, it is not of much help in answering this question. Fortunately, Massachusetts had established a state vital registration system in 1843 so that we have annual mortality information.

It turns out that 1849 was a very unusual year in terms of mortality because of the outbreak of cholera which greatly inflated death rates. Thus, whereas the expectation of life for Massachusetts males at birth was 39.4 years in 1849, in 1850 it was 46.0 years, and in 1851 it was 43.0 years. Clearly, the Jacobson Life Table exaggerated the extent of mortality because it is based on census data from an unusually unhealthy year[12].

However, earlier we argued that the mortality data from the censuses were probably under-enumerated; perhaps the exaggeration of mortality due to the cholera epidemic was compensated by the under-enumeration of the mortality data. Though this is plausible, it does not appear to be true. When we analyze the Massachusetts mortality data in more detail, it still appears that the Jacobson Life Table for 1850 greatly exaggerates the extent of mortality in that state for most years during the antebellum period[13]. Thus, we need to be extremely cautious in accepting census estimates of phenomena without first checking to see whether that year might have been atypical.

Since most studies of aggregate census data rely on the published summaries of the censuses, we also need to consider the accuracy of these published summaries. If the published summaries of the census are not an accurate tally of the individual census returns, then studies at the aggregate level will be using incorrect data. Though relatively little attention has been paid to this issue, two recent studies of the accuracy of the printed census summaries suggest that there is a serious problem for the use of at least some aggregate data.

Edward Muller has investigated the reporting of the populations of smaller towns in the federal censuses through 1870[14]. Although the populations of most towns larger than 2500 were published regularly after the 1810 census, he found that the population of smaller towns were either under-reported or the towns were entirely omitted from the published census summaries. Muller found that the pro-

[12] For a critique of the Jacobson Life Table, see Vinovskis, Maris A., The Jacobson Life Table of 1850: A Critical Re-Examination from a Massachusetts Perspective, in: Journal of Interdisciplinary History, 8, No. 4 (Spring 1978), pp. 703–724.

[13] For additional analyses of nineteenth-century mortality patterns, see Jaffe, A.J., and Lourie, W. I., An Abridged Life Table for the White Population of the United States in 1830, in: Human Biology, 14 (1942), pp. 352–371; Yasuba, Birth Rates, pp. 86–96; Thompson, Warren S., and Whelpton, P. K., Population Trends in the United States, New York 1933, pp. 228–240; Vinovskis, Maris A., Mortality Rates and Trends in Massachusetts Before 1860, in: Journal of Economic History, 32 (1972), pp. 184–213.

[14] Muller, Edward K., Town Populations in the Early United States Censuses, in: Historical Methods Newsletter, 3, No. 2 (March 1970), pp. 2–8.

portion of populations of all towns (over 100 inhabitants) in forty-one counties in southwestern Ohio and southeastern Indiana from 1820 to 1860 ranged in the published censuses from nineteen percent to seventy-three percent (the average for the period was forty-one percent). In other words, studies of small town populations based on the published nineteenth-century censuses would be very inaccurate.

The accuracy of the published mid-nineteenth-century manufacturing returns for Wisconsin have been studied by Margaret Walsh[15]. She found that in only 19.8 percent of the cases did the figures compiled from the manuscript manufacturing census coincide with those in the printed census summaries. In most instances the differences were only slight. However, in about one third of the cases at the county level, the differences between the manuscript and printed censuses were at least of the magnitude of ten percent. Fortunately, at least for the case of Wisconsin in1850 and 1860, the errors at the state level were quite small as many of the errors at the county level cancelled each other out. Thus, Walsh concluded that the major advantage of using the manuscript census of manufacturing rather than the printed summary is that the former provides more accurate information at the county level.

Both of these examples illustrate the problems of using the aggregate census returns without checking the original manuscript returns. Though for most variables we would not anticipate a very large difference between the manuscript and the printed returns, one should at least consider the possibility. Unfortunately, the logistical problems entailed in such a verification procedure are enormous — especially for studies which focus on the country as a whole rather than just one state. Nevertheless, we should at least be aware of this potential source of error and how it might affect our analyses — especially at the county level.

After we have dealt with the issues of the level of the analysis and the quality of the data, we can turn to the analysis of the materials. In many respects, the analysis of aggregate census data is the same as that of dealing with any other data sets. There are, however, some important and interesting conceptual and statistical issues raised by the way in which historians have dealt with aggregate censuses that should be discussed.

One of the most difficult problems in any analysis is to develop from the available data the appropriate indices to test our hypotheses. This is particularly difficult with nineteenth-century aggregate census data since we do not have good information for many of the socio-economic variables we would like to include in the analysis. Historians have sometimes developed indices that are inappropriate or inadequate measures of the concepts being investigated.

For example, economic historians such as Yasukichi Yasuba and others have argued that the increasing scarcity of readily available farmland in nineteenth-century America accounts for the decline in fertility among the rural white popula-

[15] Walsh, Margaret, The Census as an Accurate Source of Information: The Value of Mid-Nineteenth Century Manufacturing Returns, in: Historical Methods Newsletter, 3, No. 4 (September 1970) pp. 2–13.

tion during these years[16]. As his measure of the availability of farmland, Yasuba calculated the number of persons per 1000 arable acres. But his index was based on the cropland in 1949 and properly has been criticized for reflecting the levels of twentieth-century farming technology and practices rather than nineteenth-century agricultural potential.

The most recent effort by Forster and Tucker calculates the number of white adults per farm, using the white adult farm population in the census year under investigation and the number of farms in 1850, 1860, and 1880. Their index has the advantage of reflecting nineteenth-century farming conditions and practices more accurately than Yasuba's measure[17].

However, even Forster and Tucker's index of land availability leaves much to be desired. At the state level, an index of white adults per farm is highly correlated with the percentage of the population engaged in nonagricultural occupations and with the percentage of the population in urban areas. Therefore, we cannot be sure whether the high correlation between the white adult-farm ratio and the white refined fertility ratio is due to the availability of farms, to the percentage of the population in nonagricultural occupations, or to the percentage of the population living in urban areas.

The number of white adults per farm is not only an ambiguous measure of land availability in terms of being highly correlated with other indices, but it is also conceptually weak in that it does not reflect the relative cost of establishing a farm household. When economists speak of the availability of farms, they are in effect considering the relative costs of establishing a farm. Forster and Tucker's measure of agricultural opportunity implicitly treats all farms as equally priced, though in reality there are wide differences in the costs of farms in ante-bellum America. Thus, to take an extreme example, the average value of a farm in 1860 in Kansas was $1,179, whereas the average value of a farm in Louisiana was $11,818 in the same year. Surely it was more difficult for a young man to purchase a farm in a state such as Louisiana than in Kansas[18].

Historians have used a variety of statistical methods to analyze aggregate census data. One of the most common procedures among political historians is the use of relatively homogenous units as the basis of their analysis. Thus, communities with a high percentage of German Lutherans are used as an indicator of how German Lutherans were voting in a given election[19].

Though this method of analysis has been used extensively in American political history, it has been properly criticized for using very atypical groups of immigrants

[16] Yasuba, Birth Rates.

[17] Forster and Tucker, Economic Opportunity.

[18] Vinovskis, Determinants; Vinovskis, Maris A., Demographic History and the World Population Crises (Chester Bland-Dwight E. Lee Lectures in History), Worchester/Mass. 1976.

[19] Jensen, Richard J., The Winning of the Midwest: Social and Political Conflict, 1888–1896, Chicago 1971; Kleppner, Paul, The Cross of Culture: A Social Analysis of Midwestern Politics, 1850–1900, New York 1970.

– those who chose to or were forced to live together. There is no reason to assume that German Lutherans living in a community that was composed almost entirely of their countrymen voted the same way as those German Lutherans living in more heterogeneous communities. Furthermore, there is an implicit assumption in these studies that ethnicity or religion are the major determinants of voting behavior rather than some other variable such as occupation or wealth. By using relatively homogeneous communities as indicators of voting behavior rather than applying multivariate techniques of analysis to all of the communities in that area, these studies have under-utilized the type and variety of factors that should be considered in any mass voting analysis[20].

Another common procedure for analyzing aggregate census data is to cross-tabulate the data. Unfortunately, this method usually permits us to study only two variables at a time. If we try to use cross-tabulation techniques to control for a third or forth variable by further subdividing our data, the number of entries in our cells often becomes so small as to hinder our analysis. As a result, historians are increasingly turning to multivariate techniques such as regression analysis[21].

In most instances, multiple regression analysis is preferable to cross-tabulation in the analysis of aggregate census data. But one must be very careful in using regression analysis or any other statistical procedures not to violate the fundamental assumptions on which they are based. Historians have not always paid sufficient attention to these problems, For example, multiple regression analysis assume that the independent variables are independent of each other. Naturally, in any real life situation, the independent variables will be interrelated, but usually not at such a high level as to invalidate the analysis. There are situations where the independent variables are so highly correlated, however, that we encounter the problem of multicollinearity. There are examples in the historical literature of where scholars have included independent variables in their multiple regression analyses that were too highly correlated with each other[22].

Finally, we should briefly mention one of the major debates among scholars using aggregate census data – the issue of inferring individual characteristics from ecological data. Most historical studies using aggregate census have made inferences

[20] For a good critique of these studies, see Wright, James E., The Ethnocultural Model of Voting: A Behavioral and Historical Critique, in: American Behavioral Scientist, 16, No. 5 (May/June 1973), pp. 653–674.

[21] Some historians have tried to defend the use of cross-tabulation of data instead of employing multivariate techniques. For an interesting though somewhat misleading exchange on this issue, see Katz, Michael B., Who Went to School?, in: History of Education Quarterly, 12 (Fall 1972), pp. 432–454; Denton, Frank, and George, Peter, Socio-Economic Influences on School Attendance: A Study of a Canadian County in 1871, in: History of Education Quarterly, 14 (Summer 1974), pp. 223–232; Katz, Michael B., Reply, op. cit., pp. 233–234; Denton Frank, and George, Peter, Socio-Economic Influences on School Attendance: A Response to Professor Katz, op. cit., (Fall 1974), pp. 367–369; Calhoun, Daniel H., Letter to the Editor, op. cit., (Winter 1974), pp. 545–546.

[22] E. g. Forster and Tucker, Economic Opportunity.

only at the ecological level of the township, county, or state. But during the past five years, some historians, particularly those in the area of mass politics, have tried to infer how individuals voted on the basis of aggregate census and electoral data.

The origin of this debate goes back to W. S. Robinson's influential article on the ecological fallacy in which he warned that ecological correlations are not the same as individual level correlations[23]. Since then, several scholars such as Leo Goodman and others have tried to develop methods of using ecological regressions to estimate individual level behavior[24]. In the area of historical analysis, Allan Lichtman and Morgan Kousser have applied these techniques to the study of electoral behavior[25].

It is possible to estimate individual characteristics from ecological data under certain very limiting assumptions. Thus, if we are willing to assume that the relationship between our variables will be constant across all of our units, it is possible to make reasonable inferences about the behavior of individuals from aggregate census returns. However, it is rarely the case that the relationship between any two variables will be constant across all units – especially since there is a tendency for individuals either to move to areas with certain characteristics or to adjust their behavior in their new environment. In other words, though it is possible in certain situations to make inferences about individual level behavior from aggregate census returns, it usually presupposes the type and extent of knowledge about these variables that we simply do not have. As a result, though the use of ecological regression analysis to make estimates of individual characteristics has stimulated better efforts to specify variables and their pattern of interaction, it is unlikely that we can safely use such techniques to make firm estimates for much of the historical data available from the aggregate censuses[26].

[23] Robinson, William S., Ecological Correlation and the Behavior of Individuals, in: American Sociological Review, 15 (1950), pp. 351–357.

[24] Goodman, Leo A., Ecological Regressions and Behavior of Individuals, in: American Sociological Review, 18, No. 6 (December 1953), pp. 663–666; Goodman, Leo A., Some Alternatives to Ecological Correlation, in: American Journal of Sociology, 64, No. 5 (March 1959), pp. 610–625.

[25] Lichtman, Allan J., Correlation, Regression and the Ecological Fallacy: A Critique, in: Journal of Interdisciplinary History, 4 (1974), pp. 417–433; Lichtman, Allan J., Critical Election Theory and the Reality of American Presidential Politics, 1916–1940, in: American Historical Review, 81, No. 2 (April 1976) pp. 317–349; Kousser, J. Morgan, Ecological Regression and the Analysis of Past Politics, in: Journal of Interdisciplinary History, 4, No. 2 (Autumn 1973), pp. 237–262; Kousser, J. Morgan, The Shaping of Southern Politics: Suffrage Restriction and the Establishment of the One-Party South, 1880–1910, New Haven 1974.

[26] For further discussions of the problem of inferring individual behavior from ecological data, see Dogan, Matthei, and Rokkan, Stein (eds.), Quantitative Ecological Analysis in the Social Sciences, Cambridge/Mass. 1969; Orcutt, Guy H., et al., Data Aggregation and Information Loss, in: American Economic Review, 58, No. 4 (September 1968), pp. 773–787; Feige, Edgar L., and Watts, Harold, An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data, in: Econometrica, 40, No. 2 (March 1972), pp. 343–360; Hammond, John L., Two Sources of Error in Ecological Correlations, in: American Sociological Review, 38, No. 6 (December 1973), pp. 764–777; Wasserman, Ira M., and Segal, David R., Aggregation Effects

II. Individual Use of Census Data

Much of the recent interest in the use of the censuses has been focused on the individual level data available on the manuscript censuses. As we noted earlier, it was not until 1850 that the census began to enumerate their data on the basis of the individual rather than the household. Furthermore, the manuscript census of 1900 has not been available to scholars until very recently and then so only under conditions of restricted access in terms of where one can use them (either in Washington D. C. or in one of the regional depositories). Yet the use of individual level census information has become a major activity of American social historians today and is likely to grow in importance in the near future as historians begin to utilize these individual level returns even more effectively.

Most of the aggregate census studies have been done at the national or regional levels. Only a few have focused on just one state and even fewer have studied only a portion of a state. Studies using individual level census data, on the other hand, have almost always been done on a small geographic area – usually a town or city. In fact, most of the individual level census analyses have been done by urban historians who have studied a particular community such as Boston or New York City[27].

The almost exclusive focus of individual level analyses on single urban communities has been rather unfortunate in at least two respects. First, very little effort has been made to design research projects to include different types of urban and industrial development for comparative purposes. Second, the reliance on only urban areas has made it impossible to separate analytically the effects of urban development from more general changes within that society. In other words, by not having any rural control areas in their analyses, for example, researchers cannot be certain whether the changes experienced by any group over time within a city are the result of the impact of urbanization on their lives or the consequence of more general developments within that society as a whole.

There are alternatives to the single community focus. One such example is the study of eight Essex County (Massachusetts) communities in 1860 and 1880. This study, initiated by Tamara Hareven and myself, was designed to study the interaction between community structure and family life in one small area. By restricting the analysis to eight communities within one county, we minimized any regional differences in our analysis. We do not, of course, claim that this particular county is in any way typical or representative of the country as a whole, but only that the di-

in the Ecological Study of Presidential Voting, in: American Journal of Political Science, 17, No. 1 (February 1973), pp. 177–181; Shively, W. Phillips, ‚Ecological' Inference: The Use of Aggregate Data to Study Individuals, in: American Political Science Review, 63, No. 4 (December 1969), pp. 1183–1196.

[27] A notable exception to this approach is Blumin, Stuart M., The Urban Threshold: Growth and Change in a Nineteenth-Century American Community, Chicago 1976.

versity of activities within it permit interesting and useful analyses to be done at both the communitiy and household levels[28].

The communities selected within Essex County provided a variety of experiences and opportunities for their inhabitants. We chose three large, urban areas with different types of economic activity – Lawrence, Lynn, and Salem. Lawrence was a new city developed around the textile industry while Lynn was an old city dominated by the shoe industry. Salem was an old commercial center that only became heavily industrialized after the Civil War. In addition, we selected five rural areas – Boxford, Hamilton, Lynnfield, Topsfield, and Wenham. Though all of these communities were small in term of their population size, they also varied in their economic activities. Some, like Boxford, were almost totally agricultural, while others, like Lynnfield, had a sizable proportion of their population already engaged in the shoe industry.

We must approach the use of individual level census data within the perspective of their broader environment. The lives of women living in nineteenth-century Pittsburgh, for example, were quite different than those of women in Lawrence because of the different employment opportunities available to them in those two communities. By consciously trying to select areas of varying population size and economic activity, we can develop a more useful setting for exploring individual level census information. Studies of individuals in the past which do not even consider the impact of the nature of the community in which these people lived are badly flawed. Urban historians in particular would greatly enhance their analyses by trying to develop more controls within their research projects in order to permit them to examine the relationship between family life and community setting.

One might argue that we should forego either the single urban focus or even the broader approach suggested by the Essex County project for a national sample of households from the manuscript censuses in 1850, 1860, 1870, 1880, and 1900 (the manuscript returns for 1890 were destroyed in a fire). There is considerable merit in the idea of a national sample from the nineteenth-century censuses – particularly if the sample sizes were large enough to reflect rural-urban as well as regional differences. This approach, however, does not negate the need for identifying individuals and households within the context of the communities in which they lived since we could attach some of that information even in the national sample.

Though a national sample from the nineteenth-century manuscript censuses is a good idea and should be given very high priority in the near future, it will not eliminate the need for more indepth analyses at the local level. Local studies can deal with the interaction of family life and community opportunities in a way is difficult to study at any other level. On the other hand, future local community studies need to be designed so that they are more than just another example of a Newburyport or Boston.

[28] For studies using the Essex County data, see Hareven, Tamara K., and Vinovskis, Maris A. (eds.), Demographic Processes and Family Organization in Nineteenth-Century American Society, Princeton 1978; Hareven, Tamara K. (ed.), Family Transitions and the Life Course, New York 1978.

So far we have discussed the type of setting from which the individual level data should be gathered – ranging from a single urban community to the nation as a whole. Now we will turn to the issue of how these data should be assembled once we have selected the appropriate geographic areas for analysis.

In most studies, the unit of analysis is the household. The data on the household are assembled either by having a separate card for each member of the household or by simply summarizing some of the household characteristics from the individual returns. Though the latter method is usually much less expensive in terms of coding time and keypunching expenses, it is also more apt to produce errors and makes it very difficult to use those data for some other purpose. Thus, one of the pioneering studies from the manuscript census is Merle Curti's study of Trempealeau County (Wisconsin) which assembled its data by summarizing the household information from the census[29]. Unfortunately, Curti and his associates were not particularly interested in many of the family issues that historians today find so intriguing. As a result, it is virtually impossible, short of going back to the original manuscript census returns, to reanalyze the Trempealeau County data for such issues as the pattern of school attendance or marital fertility. If these data had been organized along individual as well as household lines, they should be of much greater value to us today.

Some of the studies based on individual census data collect information on all of the inhabitants in the area. Many of the others sample the data. Unfortunately, many of the historical studies using some form of sampling are very badly flawed either in the way the data were sampled or in the size of the sample itself. For example, some studies have drawn samples from very different sources but have treated them as comparable anyway. Stephen Thernstrom's analysis of social mobility is based on samples from five different sources – the 1880 manuscript federal census returns, Boston's 1910 marriage license registers, the 1930 Boston birth records, the Boston City Directory for 1958, and Edward Laumann's survey sample of the suburban communities of Cambridge and Brighton[30]. These are very different sources of data and are not comparable to each other. Nevertheless, Thernstrom's analysis utilizes all five rather interchangably without adequately considering the possible biases that may have been introduced.

The second major sampling problem characteristic of many of these studies is the small number of households sampled. Historians have been almost totally unaware and unconcerned about the problem of errors introduced by using samples rather than total populations. For instance, a recent study of southern Michigan drew random samples for Detroit of 70 households in 1850 and 102 households in 1880[31]. The analysts then proceeded to investigate the data for Detroit by the

[29] Curti, Merle, The Making of an American Community: A Case Study of Democracy in a Frontier County, Stanford 1959.

[30] Thernstrom, Stephen, The Other Bostonians: Poverty and Progress in the American Metropolis, 1880–1970, Cambridge/Mass. 1973.

[31] Bloomberg, Susan E., et al., A Census into Nineteenth-Century Family History: Southern

occupation of the head of the household for the native-born and the foreign-born populations. The sampling error for many of their calculations is so high as to make many of their conclusions statistically unreliable. Unfortunately, this misuse of census data by drawing samples that are much too small is a very common problem among American historians today.

Even if the sample sizes selected are adequate for the purposes of the study, one needs to decide which sampling procedure to follow. Though many of these historical studies claim to be based on a random sample of households from the manuscript census, most of them are really based on a systematic sample. That is, most of the studies have sampled every n^{th} household in the manuscript census rather than giving each household an equal chance of being selected at random. The use of a systematic sample is a defensible procedure for most purposes, but the manner in which it has been done by some historians is questionable. In order to facilitate the tracing of households in different censuses, Thernstrom eliminated any households with very common first and last names[32]. Though it is understandable why he would like to use such a procedure, it is not defensible statistically since the characteristics of individuals with very common names are not the same as those with uncommon names.

Another criticism of most, though not all, studies of urban areas is that they are not designed to investigate differences within those communities. Most of the new urban historians do not attempt to link individual level census data with their location within the city. A study of marital fertility at the household level in two Boston neighborhoods, however, has demonstrated the importance of not treating the city as a homogenous entity[33]. There are some urban historians who have tried to cope with this issue in their research. The two large-scale urban history projects that have dealt the most effectively with the problems of space and neighborhood are Theodore Hershberg's analysis of Philadelphia and Olivier Zunz's study of Detroit[34]. In both of these studies, there is a conscious effort made to study family life within the context of their local neighborhood within the city.

Since many of the issues relating to the quality of the individual level census data have already been touched on in the previous section, we will focus instead on some of the conceptual and empirical problems of using these data. One of the

Michigan, 1850–1880, in: Journal of Social History, 5 (1971), pp. 26–45.

[32] Thernstrom, The Other Bostonians.

[33] Hareven, Tamara K., and Vinovskis, Maris A., Marital Fertility, Ethnicity, and Occupation in Urban Families: An Analysis of South Boston and the South End, in: Journal of Social History, 9 (1975), pp. 69–93.

[34] For a detailed discussion of the Philadelphia Social History Project, see Hershberg, Theodore, The Philadelphia Social History Project: A Methodological History, Doctoral dissertation, Stanford University 1973; Hershberg, T., guest editor, A Special Issue: The Philadelphia Social History Project, in: Historical Methods Newsletter, 9 (March–June 1976). Zunz, Olivier, Detroit en 1880: Espace et Ségrégation, Center for Research on Social Organization Working Paper, No. 121, University of Michigan 1975.

most widely discussed and analyzed issues is the problem of measuring social mobility using the available occupational scale. The development of an occupational scale is one of the major problems in this area[35]. Nineteenth-century occupations are not identical to those in the twentieth century. For example, whereas school teachers have a relatively high status today, this was not true in the past. In fact, many female school teachers left the classroom for the factory which paid them much better wages in the ante-bellum period[36]. Though this change of jobs was not seen as downward mobility by most people in the nineteenth century, it would appear as downward mobility in most of the social mobility studies which rank occupations into professional, semiprofessional, white collar, skilled, semiskilled, and unskilled categories. Similarly, the status of shoemakers varies greatly over time since they were considered skilled artisans in Lynn in the 1830's and 1840's, but became more like factory workers after the Civil War with the introduction of new technology and organization in the shoe industry[37].

Another problem in studies using individual level data is that the occupation of the head of the family is used as the index of its status and well-being without taking into consideration the number of other wage-earners and dependents on that family. Recently, we have tried to develop a broader approach to this issue[38]. Though the occupation of the parent is a very useful and important indicator of the economic situation of the family, it is not the only economic data we would like to have. Ideally, we would measure the actual consumption needs of the family, as several contemporary studies have done. Unfortunately, such data are unavailable to us historically. We can go beyond just the occupation of the head of the household, however, by taking into consideration the number of individuals in the family who are employed as well as the number of consumers within that family.

Since the earning and consuming ability of individuals varies by age and sex, we adjusted our data by a set of weights to take these factors into consideration. Our work/consumption index is therefore a crude measure of the number of working units in each family divided by the number of consuming units. Though this index does not fully capture the individual family variations in income and consumption

[35] Many articles point to the problems in classifying occupations. See for example, Griffen, Clyde, Occupational Mobility in Nineteenth Century America: Problems and Possibilities, in: Journal of Social History, 2 (1972), pp. 310–330; Katz, Michael, Occupational Classification in History, in: Journal of Interdisciplinary History, 3 (1972), pp. 63–88; Conk, Margo Anderson, Occupational Classification in the United States Census: 1870–1940, in: Journal of Interdisciplinary History, 9, No. 1 (Summer 1978), pp. 111–130.

[36] Bernard, Richard M., and Vinovskis, Maris A., The Female School Teacher in Ante-Bellum Massachusetts, in: Journal of Social History, 10, No. 3 (Spring 1977), pp. 332–345.

[37] Dawley, Alan, Class and Community: The Industrial Revolution in Lyon, Cambridge/Mass. 1976.

[38] Kaestle, Carl F., and Vinovskis, Maris A., From Fireside to Factory: School Entry and School Leaving in Nineteenth-Century Massachussetts, in: Hareven, Tamara K. (ed.), Family Transitions and the Life Course, New York 1978; Mason, Karen, et al., Determinants of Women's Labor Force Participation in Late Nineteenth-Century America, in: op. cit.

needs, it does provide at least a beginning toward measuring a family's economic situation rather than just relying on information on the head of the household.

In our analyses of children attending school or the participation of women in the labor force, we used the work/consumption index in conjunction with the occupation of the head of the household. Though we anticipate that our particular formulation of this index may eventually be modified as researchers experiment with different weighting systems, it has proven to be useful both conceptually and empirically in our studies to date.

These are only a few of the conceptual problems involved in using individual level census data. Though I shall not produce more examples of other types of new and useful indices that can be developed from the manuscript census, such as an index of marital fertility, it should be pointed out that historians have not been particularly imaginative or aggressive in developing new ways of using the census data at the individual level. Many of the recent efforts along these lines are the result of trying to imitate as best as possible from the available census manuscripts some of the more interesting indices that have been developed by other social scientists studying the contemporary family.

Perhaps the most important advance in the use of individual level census data is the effort to use these data to estimate life course patterns. That is, historians are now starting to use the age-specific data from the manuscript censuses to reconstruct the probable life course experiences of individuals in the past. This is a particularly fruitful endeavor when we have data from more than one census and can follow different age-cohorts over time[39].

The effort to reconstruct the life course of individuals from census data is very difficult because we usually cannot follow the same individuals over time. Instead, historians have to recreate artificial cohorts of individuals based on age-specific rates of individuals in different time-periods. The problem is that the people living in Lynn in 1860 are not necessarily the same ones living there in 1880. If an area experiences considerable in- or out-migration, as most urban communities did, we may have a very distorted picture of a cohort's life experiences on the basis of cross-sectional data from a small geographic area[40]. One way of minimizing this problem (or at least of sensitizing ourselves to it) is to use the aggregate age-specific census returns in conjunction with age-specific mortality estimates to calculate the net migration in the communities which we are investigating. In this way we can at least hazard some guesses about the type of biases introduced in our analysis by the fact that we have not drawn our samples from a closed population[41]

[39] Elder, Glen H., Age Differentiation and the Life Course, in: Annual Review of Sociology, 1 (1975), pp. 165–190; Vinovskis, Maris A., From Household Size to the Life Course: Some Observations on Recent Trends in Family History, in: American Behavioral Scientist, 21, No. 2 (November/December 1977), pp. 263–287.

[40] Wells, Robert V., On the Dangers of Constructing Artificial Cohorts in Times of Rapid Social Change, in: Journal of Interdisciplinary History, 9, No. 1 (Summer 1978), pp. 103–110.

[41] Kaestle, From Fireside to Factory; Vinovskis, From Household Size to the Life Course.

Very few of the historical studies using the manuscript censuses have tried to analyze the life course of individuals. Despite the inevitable, and in some ways insoluable, methodological problems associated with such an approach, it is a much better way of organizing and operationalizing our data. A particularly useful perspective on the life course approach is provided in the writings of Glen Elder – a sociologist who has dealt extensively with family patterns and changes historically using a life course approach[42].

Finally, I will close by considering some of the statistical techniques that have been used to analyze the individual level census data. Most historians have cross-tabulated their data with the inevitable and obvious shortcomings of such a procedure. A few have even introduced the use of multiple regression analysis with dummy variables to deal with categorical variables available for individuals from the manuscript census returns.

In our own work with the individual level census data, we have found that neither cross-tabulation nor multiple regression analysis suits our needs. The cross-tabulation of data simply cannot handle the complexity of factors we want to control in our analysis. Multiple regression analysis using dummy variables is quite adequate from a statistical perspective, but it is very unsatisfactory in terms of presenting the results to other historians who are less mathematically oriented. Therefore, we have turned to multiple classification analysis (MCA) instead[43].

Multiple classification analysis is a form of multiple regression analysis with dummy variables which express results in terms of adjusted deviations from the grand mean (overall average) of the dependent variable of each of the various classes of the predictor variables. For example, MCA answers the question: how much of the likelihood of going to school was associated with being the child of an unskilled laborer, while controlling for other variables such as the age of the child, the ethnicity of the parents, and the community in which the child lived? Similarly, it provides an approximate answer to the question: *ceteris paribus*, what is the effect on youths' school attendance of the family's life course stage as measured by the age of the parents? MCA „controls" for other variables by assuming while it looks at one class of a predictor variable that the distribution of all other predictor variables will be the same in that class in the total population, thus „holding constant" their effects. Although traditional multiple regression programs also do this, MCA has three advantages: it does not require variables to be interval variables, it does not require or assume linearity and thus can capture discontinuities in the direction of the association and, finally, it is useful descriptively because it presents the reader with the gross effects of a predictor class, that is, the actual mean of each class, as well as the mean after adjusting for the influence of other variables. As historians learn to use MCA in their analyses of the individual level census data, they will find it to be a very useful way of analyzing categorical data as well as a relatively simple way of presenting their complex findings to our collagues.

[42] Elder, Glen H., Children of the Great Depression: Social Change in Life Experience, Chicago, 1974.

[43] Kaestle, From Fireside to Factory; Mason, Determinants; Rosenkrantz, Invisible Lunatics.