Chapter 2

Data Comparability Problems in the Study of the Soviet Population

ROBERT A. LEWIS

Problems in the comparability of census data—that is, the extent to which definitions and geographic units vary over time—constitute a major obstacle to research in the social sciences. The purpose of this chapter is to discuss the problems of data comparability in the Russian and Soviet censuses with respect to territory, the definition of various categories, and the periods between censuses. Also, details of a major effort to overcome many of these technical difficulties will be provided as an illustration both of the magnitude of the data comparability problems inherent in the Russian and Soviet censuses and of the procedures by which these problems can be mitigated.

The Significance of the Study of Population Processes

To emphasize the importance of census data and their use in social science research, I will preface my remarks on problems of data comparability with some thoughts on the significance of the study of population in the broader socioeconomic context. Simply put, if relevance or significance in the social sciences is defined in terms of increasing our understanding of ourselves and the society and world around us, then a knowledge of basic demographic processes is essential because these processes involve fundamental aspects of human experience and constitute a major force shaping any society. We are born, we age, we go to school, we work, we move, we reproduce, and we die. Censuses are our chief sources of data for the study of these demographic processes and their determinants and the manner in which the size and structure of a population interrelates with the socioeconomic and natural environments.
The most momentous and pivotal events in human history have been the dramatic declines in mortality and fertility and the rapid urbanization of society which have occurred in the West in the past century or so. These profound changes have resulted in other pervasive societal trends, affecting virtually all aspects of society. It is probably fair to say that these events have not been adequately reflected in historical writings.

In addition to economic, political, geographic, and social factors, demographic forces shape a society and must be studied if we are to understand a society and make guarded forecasts as to future trends. With respect to forecasts, more can be said about demographic processes than about other social or economic trends because the workers and parents for the next twenty years have already been born. The principal reason why a knowledge of demographic processes is essential to understanding social change—aside from the obvious fact that they are interrelated—is that so many crucial events affecting society are age specific, including birth, death, marriage, migration, work-force participation, educational attainment, crime, consumption, voting, retirement, and generational differences in life-styles. Consequently, major shifts in age distribution resulting from demographic trends can have desirable or undesirable effects on a society. For example, the impact of the rise in fertility which resulted in the “baby boom” generation in the United States on nurseries, schools, colleges, the work force, the housing market, and eventually the retired population and the social security system is well known. To cite another case, the decline in death rates in developing countries brought about rapid population growth and a younger population, which in turn resulted in a less favorable ratio of productive to nonproductive individuals. Under certain conditions, this demographic situation affected savings and investment and hindered social and economic development; these and other problems that rapid population growth has caused in the developing world are well documented.

Changing age composition can also have beneficial effects on a society. Thus, a decline in fertility—which occurred in most developed countries—results in a declining rate of growth of new entrants to the work force, a development that over time should alleviate unemployment problems (if work-force participation rates and economic conditions remain constant). Fertility decline also results in an increasing share of the population in the working ages and thus a decreasing dependency burden. Other alterations in the structure of a population, such as the sex ratio or ethnic composition, can also affect a society.

Suffice it to say, then, that virtually every social or economic problem facing a country has a demographic dimension. However, one must avoid what might be termed “demographic determinism” in appraising the effects of population change on a society because demographic
processes interrelate with a variety of other forces; after all, the "sins" of environmental and economic determinism resulted from exaggerating the influence of one factor in the human milieu. One illustration of single-factor determinism in the Soviet context is the purported psychological and political impact of the decline in the Russian ethnic group's share of the Soviet population. Clearly, dominance in a multiethnic state does not relate solely to brute numbers but to political, historical, economic, cultural, and ideological factors as well. Moreover, there is no indication that the dominance of the Russians has varied relative to their share of the population. In the Russian Empire in 1897, the Russians comprised 44 percent of the population, a figure that rose to 53 percent in 1926—primarily because of the loss of territory inhabited by non-Russians—and to 58 percent in 1939. Largely owing to the territorial acquisitions during World War II, the percentage of Russians in the Soviet population declined to 55 percent in 1959 and, because of declining natural increase compared to other Soviet ethnic groups, to 53 percent in 1970 and 52 percent in 1979. There is no reason to believe that political or psychological trauma or a diminution in the dominance of the Russians will ensue when their share slowly drops below 50 percent. Likewise, the changing ethnic composition of draftees into the Soviet armed forces or the decline in the growth of the work force does not necessarily herald demographic crises. The essential point is that population changes, important as they may be, are not the only factors to be considered and that the effects of demographic trends on the Soviet society and polity should not be overstated if we are to assess conditions in that country.

The Importance of Data in Social Science Research

Data, theory, and method are the essential elements of research. Scholarly research requires a rough balance in these elements in order to avoid the common academic inclination toward excessive and tedious description, method for its own sake, and theory unconfirmed and unrelated to reality. Whereas theory and method are the subject of considerable academic attention and admiration, data considerations are generally neglected and considered mundane. Even in demography, a heavily empirical discipline, little has been written on data problems relative to work on theory and method. Most failures in research, however, can be attributed to lack of data. If theories are lacking, new hypotheses can usually be formulated, and if new methods are required, they can ordinarily be devised, but as a rule appropriate data cannot be generated.
Thus, the first concern in applied social science research is acquiring the requisite data; the application of method and theory is normally of secondary importance. Because data are not collected specifically for one's own research and are frequently lacking, in most instances it is necessary to use surrogates or to employ a complex methodology to deal with less than satisfactory data. On the other hand, if data are adequate, elaborate methods are often not required. The above, of course, is not to deny an appropriate respect for theory and method.

In the social sciences, data skepticism is of the utmost importance and is the mark of scholarly maturity. The mindless acceptance of published data as accurate or representative of the phenomena under investigation often results in erroneous interpretation and bad scholarship. Kingsley Davis has rightly deplored the tendency in demographic research to apply sophisticated methods to unevaluated and often unreliable data without a conceptual framework and then uncritically accept the results (Davis, 1965: 147). If the essence of good scholarship is asking the right questions, then the questions that are asked should be appropriate to the accuracy of the data at hand. Thus, investigating operational definitions and evaluating data are crucial steps in research. Some questions can be answered with relatively bad data, but others cannot. Instead of asking if the data are accurate, the researcher should consider how accurate they are and if they are sufficiently representative of the subject under investigation.

But above all, with respect to data skepticism, the researcher should realize that virtually all population data contain errors resulting from deficiencies in collection and tabulation and from inconsistencies in response to questions. These types of errors occur in censuses, even though census data are usually collected more systematically and result from more straightforward questions than most other kinds of socioeconomic data. Census questionnaires, instructions to census takers, and census sample surveys are of great importance when evaluating the accuracy of the data. Errors in the registration of vital statistics are even more common because of the difficulty in defining some demographic events (such as what constitutes a birth), regional variations in the use of medical facilities, and other factors that affect reporting.

In addition to its crucial importance in analysis, the disaggregation of data is very useful in evaluating accuracy because there are frequently geographic differences in the quality of enumeration. This is especially true of the USSR. Thus, if unexpected variations in data are discovered, the first step should be to question the accuracy of the figures themselves.

Data considerations are particularly important in the study of population because the essential elements of demographic research are (1) the ordering of the data into consistent definitional and territorial categories;
(2) the evaluation of the data; (3) the calculation of standardized measures; and (4) the application of demographic and statistical methods and theory to the data for estimation and analysis. Thus, data comparability and data considerations are central in demographic research.

The Problem of Data Comparability

The basis of the general problem of data comparability is that most data are artifacts, and thus there are few uniquely correct definitions, territorial units, or temporal periods for which data can be collected. This is so because most statistical data are continuous, and their demarcation into categories is, in most instances, arbitrary (Jaffe, 1982: 5–8). In practice, one encounters three types of comparability problems: definitional, territorial, and temporal. In the first instance, if there were uniquely correct, universal, and constant definitions for demographic phenomena, there would be no problems of definitional comparability. Instead, of course, operational definitions of seemingly unambiguous events differ markedly from country to country and over time because changing conditions require changing definitions.

Likewise, the territorial units in which data are collected and published are also artifacts and can bias data. Regionalization is merely the spatial aspect of the classification problem; just as there is no all-purpose statistical or taxonomic interval or class, there is no perfect set of regions for any country. A different set of regions might be required for each aspect of population that is to be studied. In this regard, regions should be homogeneous in terms of the problem under consideration and based on one factor or a combination of factors. Also, spatial units of analysis should be delimited on a consistent basis in a way that maximizes external and minimizes internal variation. Unfortunately, population and other data are not collected at the required scale in the USSR (or most other countries) to permit the construction of an ideal set of regions for every analytic purpose. Furthermore, when comparisons are made over a long period of time, the problem of establishing a set of homogeneous regions becomes virtually insoluble if significant economic and social change has occurred. Consequently, researchers are usually forced by circumstances to use some set of existing units, but in all cases they should be aware of the set’s limitations.

A related problem to that of territorial comparability is that of scale, which is in itself extremely important in spatial analysis because results often differ depending on the scale of the units used. For example, data by union republic in the USSR are not suitable for use in most research projects owing to the differences in size as well as in population and
Data Comparability Problems

socioeconomic diversity among them; the RSFSR is almost twice the size of the United States, whereas Armenia is only about the size of Belgium, and the population of the RSFSR is about ninety-two times that of Estonia. A major data problem in the study of the USSR is that for most subjects more data are available by republic than by the more useful lower-order units.

Finally, for most research problems the time intervals between censuses should be constant. Ideally, censuses should be taken every five or ten years, so that trends and projections can be more easily measured by aging or tracking five-year age cohorts from one census to another. In many instances, however, this does not occur, and difficulties of temporal comparability arise. The steps necessary to overcome these inconsistencies and the biases that are introduced are often formidable.

The Problem of Data Comparability in the Study of the USSR

Data comparability—definitional, territorial, and temporal—is the central statistical problem in using the censuses in the study of the population and society of the USSR. Owing to severe comparability problems, trends over time which are crucial to demographic analysis cannot be established among censuses unless data are in some way rendered comparable. Largely because of these statistical obstacles, researchers have used the censuses relatively little in the study of Russian and Soviet society. These problems may be compounded because at this writing it appears that only very limited results, and virtually none of the vital age data, will be published from the 1979 census. This dramatic curtailment of census publishing in the USSR will limit the number of variables for which comparable data in all censuses can be derived. At a time when major international efforts are being made to provide basic demographic data for the world, it is indeed unfortunate that only such a limited coverage of the largest and one of the most important countries will be available.

Because the national territory of the USSR has changed several times over the last century, its present-day borders differ from those of 1897, 1926, and 1939. Consequently, to provide even comparable aggregate data based on the territory of the USSR today, researchers must gather data from East European and other censuses for the border areas formerly outside, but currently within, the USSR. However, aggregate data for a country as large and diverse as the USSR are not especially meaningful, and therefore some subnational regionalization is necessary to analyze population and socioeconomic trends. Unfortunately, frequent
and drastic changes in the internal political-administrative divisions into which census data and other data are ordered make it very difficult to compare regional, demographic, and socioeconomic characteristics over time. For example, the number of major enumeration units in 1897 was 89; in 1926, 189; and in 1959, 1970, and 1979, about 140. Furthermore, no enumeration unit in any census had the same boundaries in all census years. Thus, the major tasks with respect to territorial comparability of Soviet census data are to gather data for border areas from the censuses of neighboring countries and to reorder the Russian and Soviet data into a consistent set of territorial units.

Problems of definitional comparability—most of which are discussed in detail by others in this volume—are especially nettlesome in demographic research on Russia and the USSR. First, and most obviously, many important categories are defined differently from census to census, including such key socioeconomic indicators as urban residence, educational attainment, ethnicity, and occupation. In fact, sex and age were the only two characteristics for which the definitions were directly comparable in all censuses, although the categories in which age data were given changed from census to census. A second and more subtle problem is that little systematic attention is given to operational definitions, and consequently changes in categories are often difficult to determine. Thus, one finds confusion over apparently straightforward details, such as the number of persons enumerated at the national or subnational level; there are significant differences between the de facto (nalichnoye) and de jure (postoyannoye) populations in the Soviet censuses (Anderson and Silver, 1985).

Still another problem is that of temporal comparability, a difficulty that arises because of the irregular intervals over which censuses have been taken in Russia and the USSR. That the censuses were not conducted every five or ten years greatly complicates research on the population of the USSR. Therefore, when measuring change among the various censuses, researchers must use average annual rates of change to standardize for the different intercensal periods. They should also avoid assuming that change between two censuses was linear. In the two intercensal periods that included wars, for example, most demographic trends were not linear but fluctuated considerably. Even between 1926 and 1939, when there was no war, it is not reasonable to assume that change in such variables as mortality and fertility was linear because famine, drastic economic reorganization, and variations in socioeconomic conditions clearly had a specific temporal impact.

The 1897 census was conducted on February 9 (January 28 according to the Julian calendar), the 1926 census on December 17, the 1939 census on January 17, the 1959 and 1970 censuses on January 15, and
the 1979 census on January 17. In the computation of average annual rates, a period of 19.9 years is used for the 1897–1926 period, 12.1 years for 1926–39, 20.0 for 1939–59, 11.0 years for 1959–70, 9.0 for 1970–79, and 82.0 years for 1897–1979.

A Method to Overcome Data Comparability Problems

For the past two decades, my colleagues and I have done considerable research on problems of territorial and definitional comparability related to the Russian and Soviet censuses. As a result of these efforts, we have generated a wide array of variables that are comparable in terms of both territory and definition. In this section I will describe in brief the procedures that we derived in an attempt to solve the many comparability problems associated with the original census data (Lewis, Rowland, and Clem, 1976: 29–60; Leasure and Lewis, 1966: v–41).

As was previously mentioned, there is no ideal set of regions for the study of demographic change. Nevertheless, a careful selection of regions into which population data will be ordered is important to avoid biasing the data excessively. For our study, the nineteen major (krupnyye) Soviet economic regions of 1961 were selected as the consistent territorial units into which the original data would be ordered. We chose these regions because their scale was appropriate for regional analysis, because it was easier to order data into these regions than the more numerous lower-order political-administrative units, and because these regions were largely congruent with the ethnically based federal structure of the USSR. Moreover, the 1959, 1970, and 1979 political-administrative and census units conformed to these regions without major adjustments, and other data pertinent to demographic analysis had been presented in these regions.

In order to allocate original data into the comparable territorial units (i.e., the nineteen economic regions), we assumed that the rural population was evenly distributed within each administrative unit. Thus, a variation in area would result in a proportionate variation in rural population. For our purposes, we defined the rural population as the population not living in centers of 15,000 and over; because the definitions of “urban” varied considerably from census to census, we chose the size criterion of 15,000 (the smallest urban center for which data were available in all censuses). The procedures for allocating the rural population were simple and straightforward. The first major step was to superimpose a map of the economic regions over a map of the political units for each census year and then to measure with a polar planimeter the percentage of the area of a given political unit that fell into a given
region, a percentage that we termed the "area allocation." We used equal area maps of relatively small scale and where possible checked the results with maps of larger scale. In practice, many political units fell totally within the larger economic regions, obviating the need for allocation. Thus, the rural populations of the units that constituted an economic region were multiplied by their area allocation and summed to obtain the estimate for the rural population of that region.

The allocation of the urban population required fewer calculations because the urban population is spatially concentrated and its location known. Thus, urban centers could be allocated directly into the appropriate economic region. Accordingly, the summation of the population in centers of 15,000 or more from the given political units allocated to a given economic region yielded the urban population estimate for that region. An estimate was made of the urban and rural populations and of the total (i.e., urban plus rural) population of a given political unit which were allocated to a given region. The percentage that this allocated population represented of the total population of the original unit was termed the "population allocation." Once the total, urban, and rural populations of each unit had been apportioned to the appropriate regions, it was possible to estimate these populations for each economic region simply by a summation of all constituent allocations. These procedures—with adjustments for census dates—were also applied to border areas currently within the USSR which were outside the boundaries of the Russian Empire in 1897 or the USSR in 1926 and thus not included in the Russian and Soviet censuses.

To estimate demographic variables for the total population of each region in 1897 and 1926, we multiplied the number of persons in a given unit with certain characteristics by the population allocation. For the few border areas for which data were not available, we assumed that the characteristics of the population for which we had data could be applied to the entire population of the region or we made independent estimates, as was frequently necessary. The allocation of the demographic characteristics of the urban population was somewhat more complicated. The most logical procedure was to allocate the urban (census definition) characteristics on the basis of the urban (census definition) population, which is the percentage of the urban (census definition) population of a unit which was fitted into an economic region. The summation of the allocated figures for each region resulted in an estimate of the characteristics of the urban (census definition) population of each region. To derive an estimate of the characteristics of the urban (15,000 and over) population—our comparable urban definition—we adjusted these data by assuming that the characteristics of the urban (census definition) population could be applied in the same proportion to the urban (15,000 and
Data Comparability Problems

over) population. This assumption was not unwarranted because in fact the vast majority of the urban (census definition) population is in centers of 15,000 or more. For example, if 20 percent of the urban (census definition) population were in the 0–9 age group in a given economic region, this same percentage would be applied to the urban (15,000 and over) population to derive an estimate of this age group according to our comparable definition of urban. After the demographic characteristics of the total and urban (15,000 and over) populations of each region were derived, the characteristics of the rural population could be easily figured by subtracting the number involved in the urban (15,000 and over) population from that of the total population.

Available tests indicate that the error involved in these procedures was small. One check on the accuracy of the method is a comparison of the official area of the 1961 economic regions with the sum of the estimated area of the component units of these regions for 1897 and 1926. In most instances, this comparison yielded an error of less than 2 percent, and only once did it exceed 4 percent. These results are reassuring when we consider that in this period some parts of the country were not accurately mapped and it was necessary to transfer projections to obtain an equal area map for 1897.

Another check is a comparison of our figures with the few available Soviet estimates. For example, our estimate of the total population in 1897 within present-day borders was only about 400,000 above the official Soviet figure. Differing estimates of the population of Khiva and Bukhara probably account for most of this discrepancy, although no indication has been given as to how the Soviet figure was derived. Official Soviet estimates for the population of the Transcaucasus, Central Asia, and Kazakhstan in 1926 have also been published. Because the external boundaries of the Transcaucasus and Central Asia have changed only slightly, that our estimate and the Soviet figure are very close does not constitute a good check. However, a comparison with the Kazakh Republic (which is also an economic region) would be an excellent validation because a more complicated division of units and an uneven population distribution (related to the diversity of the natural environment) increase the chance for error in this case. Nevertheless, the difference between our estimate of the population of the Kazakh Republic in 1926 and the Soviet estimate was only about 2 percent.

Recently, the Soviet government published population estimates for the 1926 population in 1970 units at the oblast scale; typically, no indication was given as to how these figures were derived (USSR, 1975: 14–25). However, no data were provided for areas that were outside the USSR in 1926, so the western border regions and the Far East cannot be compared with our estimates (which do include these areas). Aggregat-
ing the Soviet data into the 1961 economic regions reveals a very close correspondence with our figures; in most cases they were within 1 percent of each other and never above 5 percent.

Moreover, Ralph S. Clem independently developed estimates of the 1926 population in 1959 lower-order units (oblast scale) by using the allocation procedure described above; his figures can be aggregated into the 1961 economic regions and compared with our estimate (Clem, 1977: 599–602). These two sets of estimates were also very close, almost always falling within 2 percent and in most cases within 0.5 percent of each other. All told, therefore, we have considerable confidence in the estimates that we have derived.

Summary

The most basic aspect of demographic research is the determination of trends over time, which in turn are essential in the analysis of the determinants of population change and demographic comparisons over time and space. As we have seen, population trends cannot be derived from the Russian and Soviet censuses unless major problems in territorial and definitional comparability are solved. The procedures that we derived make it possible for the first time to establish these trends longitudinally for the present-day territory of the USSR for a wide variety of census variables, such as ethnicity, urbanization, labor force, sex, age, fertility, and literacy, for the total, urban, and rural populations.

Aside from making possible the measurement of trends over time, solving problems of data comparability by these or other methods presents other advantages. Data can be put into comparable categories that permit the application of estimation procedures or that can be used to adjust other data. For example, after comparable data by five-year cohorts have been ordered into territorially consistent regions for the various census years, stable population procedures, life tables procedures, and indirect standardization can be used to estimate fertility, mortality, and natural increase. Five-year age groups can also be used for direct standardization of other variables (fertility, mortality, marriage, etc.) to take into account the effects of changing age distribution. These data can also be aggregated to make comparisons over time in age groups related to the life cycle: working-age, dependent, and retired populations.

Moreover, comparable data can facilitate comparative research, although problems of data comparability at the international scale are staggering. Comparative research can provide insights into how universal the processes under investigation are and—if they are universal—provide cautious predictions as to future trends. Also, the comparative method
Data Comparability Problems

can enable researchers to hold constant the effects of a given political or economic system. Clearly, the demographic or social processes in the USSR cannot be understood solely by studying that country, for a broader understanding of these processes is required. Therefore, whenever possible researchers should attempt to order their comparable data into categories that permit international comparisons.

The development of procedures to solve the many problems of comparability is particularly important in the Soviet context because the available population censuses constitute a major source of socioeconomic data for the study of Soviet society. Moreover, because demographic processes are an important force shaping Soviet society—or any society—and are interrelated with many other socioeconomic phenomena, a contribution to our knowledge of the population of the USSR should greatly increase our understanding of that country and facilitate our making some guarded and objective forecasts as to the course of events in the near future.

References