INTEGRATING SURVEY, DEMOGRAPHIC, AND MODELING METHODS

Howard Hogan[1], William Bell, Lynn Weidman, U.S. Census Bureau,
and Allen Schirm, Mathematica Policy Research, Inc.
Howard Hogan, ESMPD, U.S. Census Bureau, Washington, DC 20233-6200

KEY WORDS: Coverage Measurement, Intercensal Population Estimates

1. Introduction — What Are the Paradigms of Demography and Statistics?

In 1662, when John Graunt published his "Observations upon the Bills of Mortality," statistics and demography were fully integrated. Demography was the most quantitative, not only of the social sciences, but perhaps of all sciences. Statistics concerned itself with the affairs of state, of which the size and growth of the population was a central concern. However, since that time, the two have grown apart.

The disciplines of demography and statistics are now broad with diffuse borders. Each has a core body of knowledge and many areas of specialization and application. Our focus in this paper is on applications where both survey based statistical estimates and estimates produced using demographic methods are available, and thus both have information that ideally could and should be used. We also discuss several specific attempts to integrate statistics and demography in order to highlight some of the fundamental difficulties and, hopefully, stimulate further inquiry.

To begin, we need to define what we mean by "the statistical and demographic approaches" as clearly as we can. We consider demography first, then statistics.

Demography, at its broadest, is the quantitative description of any human population or population process, especially population size, mortality, fertility or migration. However, there is clearly a set of methods that might be termed demographic estimators. At its core, demographic estimation applies the demographic balancing equation, which relates population change to births, deaths, and migration. Demographic estimation can also apply mathematical models of fecundability, mortality, and stable population theory. In applying such models or other approaches, demographic estimation often utilizes certain core biological principles, such as age patterns of mortality and fertility. It can also apply certain sociological generalities. An example would be the application of common, although not universal, patterns of age and sex composition when estimating internal and external migration.

In addition, a demographer might bring to bear a wide body of knowledge of economics, sociology, history and, indeed, patterns of response and coding error. (See for example Coale and Stephan, 1962.) Much of applied demography involves the ability to gather and synthesize a wide range of information sources into a coherent analysis of population processes. Demographic estimation, then, can be seen as a synthesis, within the constraints of the core models, of available data. The data are often from different sources and of different quality. The demographer thus seeks estimates that are consistent with both the core models and the observed data.

Although statistical analysis is a broad field that may be largely defined as the application of probability theory to data analysis, we are concerned here primarily with survey estimation and analysis of survey data, an area where statistics often intersects with demography. Statisticians engaged in survey estimation are concerned with producing both point estimates and associated variances. These two are combined to produce confidence intervals – probabilistic statements about the true, underlying quantities of interest. While statistics lacks a subject matter discipline, it finds broad application across many fields through the combination of general statistical models (e.g., linear regression models) with subject matter knowledge. The latter may be expressed through a formal mathematical model, which then is analyzed statistically, or it may serve to define the particular form of the general statistical model that is used (e.g., specification of regression variables).

To speak very broadly, two primary features of statistical analyses are a reliance on formal

[1] This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The opinions expressed are those of the authors and do not necessarily reflect those of the Census Bureau or Mathematica Policy Research.

mathematical models and use of probability theory to quantify uncertainty. Statistics thus contrasts with demography which relies more on implicit models, and which also tends to treat uncertainty arising from, e.g., errors in data sources, more informally, if at all.

The differences in approach between statistics and demography can lead to conflicts. For example, statisticians might criticize demographers for producing (population) estimates without also providing objective measures of uncertainty for the estimates. However, some of the data sources used to produce population estimates have little or no error (e.g., birth and death registration data), while others (e.g., related to estimation of undocumented immigration, emigration, and internal migration) have substantial error that is not easily assessed by conventional statistical techniques. These issues concern more what a statistician would consider bias than variance, and measuring bias can be a difficult task.

One problem with integrating demographic and statistical methods is the problem of continuity. Improvements to demographic estimates can come in huge leaps. These discontinuities are often not the direct result of bringing in new data, but rather from a re-analysis of existing data. This re-analysis, however, may be inspired by a newly perceived problem, the perception being inspired by new data. For example, when new census data become available, the implications of previous demographic assumptions are made evident, so that alternative assumptions are explored. These are discussed in further detail in the following section. For now, we simply note that the leaps make modeling uncertainty difficult. Everything is stable until one day the evidence appears (or is recognized) and there is a large change.

The next three sections discuss areas where efforts have been made to combine demographic, survey, and statistical modeling methods: census coverage measurement, postcensal population estimation, and population forecasting and projection. We hope to show not only the particular aspects of each application, but also the common issues that make the integration of demographic and statistical methods difficult. We then touch upon the possible use of Bayesian analysis to address the integration problem. We conclude with some suggestions for future research.

2. Census Coverage Measurement

Starting with the 1950 Census, two methods – one statistical and one demographic – have been used to measure the coverage (undercount or overcount) of the U.S. Census. The statistical method, the post-enumeration survey (PES), is based on a sample of the population that is matched to the census records to measure the omission rate. The PES also verifies a sample of Census records in order to measure the rate of erroneous enumerations. Together, they are used to produce a measure of net census coverage error via dual system estimation (DSE). (See Anderson, 2000.)

The other method, demographic analysis (DA), produces an alternative estimate of the population that is compared to the census totals to measure net undercount. For the 1950 Census reliable birth records existed and could be used via DA to obtain population estimates only for the population under aged 15. For the rest of the population DA relied on an analysis of previous censuses, using such tools as stable population theory and knowledge about patterns of age mis-reporting. With each census, new data sources and new methods of analysis have been developed. Thus, the method of demographic analysis used for Census 2000 bears little resemblance to that used for the 1950 Census. (See Anderson, 2000.)

PES estimates have the advantage that they can be produced for sub-national areas and groups, e.g. states and ethnic groups. A disadvantage is the inability of the PES to completely measure the undercount for some hard to enumerate groups, especially adult black males and undocumented aliens. In contrast, the DA approach has the advantage of producing better estimates for some groups, including adult black males and children, but one disadvantage has been the inability to produce sub-national estimates or estimates for important ethnic groups such as Hispanics and Asians.

Even with the 1950 Census, the potential gains from exploiting the methods complementary strengths by combining estimates was evident. In one of the earliest demographic analyses of census coverage, Coale and Zelnick (1963) benchmarked their demographic estimates to census population totals as corrected by the 1950 PES. However, this benchmarking was to the estimate for females aged 15 to 29, not total population. The existence of two somewhat contradictory estimates of census coverage caused little problem so long as the only use was to generally inform data users and census planners.

By the time of the 1980 Census planning, the Census Bureau and the broader government and user community were faced with the issue of possibly correcting the census results for measured undercount for at least some official uses. This possibility quickly brought to the forefront the issue of combining the two estimation approaches to produce a best estimate. Combining the two seemed especially attractive at the time because there seemed to be a possibility of producing state level demographic analysis estimates (Siegel, Passel & Robinson, 1977.)

Around the 1980 Census, an early attempt to combine the two approaches used a traditional statistical approach. The two methods could be linearly combined using some measure of their relative precision. PES estimates came ‘equipped’ with estimates of their sampling error variances (though not with measures of their non-sampling error). The problem was then to quantify the uncertainty in the DA estimates.

Using models and approximations based on the Pearson-Tukey approximation (1965), the approach taken was to establish subjective 90 percent probability ranges for the DA estimates. However, several problems arose. First, because of the natural tendency to underestimate the variability and uncertainty in any system, there was no doubt some underestimation of DA uncertainty. Nonetheless, this could be considered no more severe than the problem of ignoring non-sampling errors in the PES estimates. With further analysis better intervals could have been established.

A second, more difficult problem was how to deal with correlations between components. A principal strength of demographic analysis comes from the required balancing of different components. Demographers simultaneously analyze all the components of a population system to fit them into a coherent whole. Age patterns must make sense, sex ratios must be within limits, etc. These constraints necessarily imply correlations between the errors in the estimates of the individual components. Unfortunately, no satisfactory method was developed to measure correlation between the components.

An equally difficult problem came from the discontinuities in demographic analysis. Changes in the data, procedures, and assumptions used to derive the estimates often come in huge leaps.[2] These changes included:

- The use of Medicare data to estimate the population aged 65 and over, beginning with the 1970 census. Indirect estimation methods had previously been used and were subject to more error than the Medicare-based estimates, especially for Blacks (Siegel, 1974, pp. 15-18.)

- The change with the 1980 Census of assuming 3 million, rather than zero, illegal immigrants during the 1970s. (See Fay et al. 1988.)

- The shift for the 1980 Census DA evaluation from Coale and Zelnick estimates to Whelpton and Campbell estimates for pre-1935 white births, and the shift from indirect estimates to Coale and Rives estimates for pre-1935 Black births. (See Fay et al. 1988, Chapter 2 and the references given there.)

- The use for the 1990 DA evaluation of a new time series of Black births, with major revisions in the assumed birth registration completeness for 1940 (lowering Black births from 1935 to 1950), and the reclassification of Black births on a "Father rule" basis (also lowering Black births, with greatest effect on births after 1975). (See Robinson et al, 1993, pp. 1063-1064 and references given there.)

- The shift with the 2000 Census from assuming incomplete registration of recent births to complete registration, affecting births from 1985 to 2000. (See Robinson, 2002, pp. 5-6 and the references given there.)

- The major revision in the 2000 DA evaluation of the components of net international migration, especially the estimates of unauthorized immigrants. (See Robinson, 2001.)

Each of these revisions was the result of careful analysis, and each almost certainly improved the estimates. As Table 1 shows, the effect of the revisions on DA levels can be appreciable (compare Rows 4 and 5 for 1980 and rows 9 and 10 for 2000). However, the effects of revisions on the age, sex, and race patterns of coverage are less – the male/female differentials remain about the same

[2] The authors gratefully acknowledge J. Gregory Robinson for providing them with this summary.

across revisions within a census and all the sets document the Black/Non-black undercount differential (Table 2).

Statistical estimates of census coverage are not immune from significant revisions when problems are discovered. For example, Accuracy and Coverage Evaluation (A.C.E.) estimates of 2000 census coverage were revised from an estimated undercount of about 1.2 percent (March 2001 estimates) to an estimated overcount of about .5 percent for A.C.E. Revision II (U.S. Census Bureau 2003). This was a larger overall revision than was made to the corresponding DA estimates. It reflected significant corrections to the underlying survey data to deal with undetected census duplications and coding errors, as well as an adjustment for correlation bias (discussed subsequently). All these adjustments addressed certain types of nonsampling errors in the March 2001 estimates, thus these errors were not reflected in the corresponding sampling variance estimates that were produced. In fact, discontinuities like the revisions noted in the DA and A.C.E. estimates do not fit well with standard statistical measures of uncertainty such as sampling variances – they can be thought of more as relating to biases than to variance. We could say that virtually all the errors in the DA estimates of population are nonsampling errors (biases) for which variances are not readily available. Nonsampling errors in census coverage survey estimates, as with any survey estimates, limit their accuracy and also limit the usefulness of conventional sampling variances as measures of estimation error. One might note that there is some potential for addressing nonsampling errors in the survey estimates by attempting to design the survey (and census) procedures and estimation methods to minimize these errors. Nonsampling errors in the DA estimates, which come from errors in administrative data sources and from reliance on assumptions where data are limited (e.g., for undocumented migration), may be less controllable.

The difficulties with quantifying uncertainty in the DA estimates led to abandoning the initial effort to combine them with statistical estimates of 1980 census coverage. Work on quantifying the uncertainty in the DA estimates continued, however, resulting in subjective measures of uncertainty in DA estimates of 1990 Census coverage (Robinson et.al, 1993). However, a more fruitful approach for combining the DA and coverage survey estimators developed from the idea of taking just the results believed to be most accurately estimated by DA, treating these as "truth" (measured without error), and using these results as controls in the statistical estimation. As noted above, while errors in the level of the demographic estimates were of concern, coverage patterns from DA are viewed as more accurate and have remained relatively stable as estimates have been revised. Most stable of all have been the DA estimated sex ratios. These clearly point to a deficiency in the coverage survey DSEs for the adult male population, especially for African-American males. Wolter (1990) suggested modifying national level male DSEs to force agreement with the sex ratios from DA for age-race groups. Bell (1993) extended the idea by using these national level estimates for males as controls and showing how alternative modeling assumptions can be used to modify subnational DSEs for males, keeping them consistent with the national level controls to address correlation bias in the male DSEs. The different proposed models all produce the same national estimates for males by age and race (Black and Non-black) that are obtained by multiplying the national level estimates for females from the DSEs by the DA sex-ratios. The different modeling assumptions produce different subnational estimates and different estimates for race groups more detailed than Black and Non-black. For example, since DA does not provide a separate estimated sex-ratio for Hispanics; the estimates for Hispanics from this approach are affected by the modeling assumption. Elliott and Little (2000) further developed this approach, casting it in a Bayesian context, though retaining the same basic modifications to the male DSEs as in the approach of Bell (1993).

3. Intercensal Population Estimates

The Intercensal Population Estimates Program of the U.S. Census Bureau develops and disseminates annual estimates of the population and its age, sex, race, and Hispanic origin characteristics for the nation, states, counties and other functioning governmental units. Traditionally, the national estimates have been based on the previous decennial census and annual estimates of components of change due to births, deaths, net international migration, and net movement of U.S. Armed Forces and civilian citizens to the U.S. developed from administrative records. State and county estimates are produced using the same components of change plus estimates of net migration within the country.

In 2002 new data sources – the Census 2000 Supplementary Survey and the 2001 Supplementary Survey – were used to improve the national estimates of change due to international migration.

This is the first step in a plan to make increasing use of survey data, especially from the American Community Survey (ACS), for improving estimates of the components of change. Previously, administrative records were the only data sources used to develop the estimates.

The following sections present a brief overview of the methodology used to obtain the vintage 2001 national population estimates for the months of calendar year 2001.

The base population is the enumeration of the resident population from Census 2000. These estimates were then updated using the following estimated components of population change for each age, sex, race, and Hispanic origin group. The following change components were calculated for each sex, race, and Hispanic origin group.

1. Births to U.S. resident women
2. Deaths to U.S. residents
3. Net international migration (legal immigration to the U.S.; emigration of foreign-born and native people from the U.S.; net movement between the U.S. and Puerto Rico; and estimates of the net residual foreign born population, including unauthorized immigrants.)
4. Net movement of U.S. Armed Forces and civilian citizens to the U. S.

The net migration of U.S. citizens not affiliated with the federal government, temporary movement of students, scholars, and embassy personnel are tacitly assumed to be zero.

Each of the data sources used to estimate these components fails to provide at least one of the characteristics – age, sex, race, and Hispanic origin – at the level of detail required for direct use in the estimates. Through the use of demographic assumptions and modeling that varies by data source, this level of detail is derived for each component. See U.S. Census Bureau (2001) for details.

An additional component of change for states and counties is migration within the U.S. The sources of data used as a basis for estimating the total population change of this component by county are: (a) consecutive year IRS tax returns for the population under age 65, and (b) Medicare enrollment records for the population age 65 and older. State total population estimates are a sum of the county estimates. State and county migration rates by age, sex, race, and Hispanic origin are based on matching consecutive year IRS tax returns and comparing the addresses of the matches. Tax returns cover about 80 percent of the population, and it is assumed that the remainder of the population follows the same migration pattern.

Age, sex, race, and Hispanic origin information about the filer(s) of a tax return is obtained from Social Security Administration files by matching on SSN. The characteristics for the dependents claimed on the return are derived, where race and Hispanic origin are assumed to be the same as the filer. (See Sink, 1999.)

The demographic population estimates are used as controls in the estimation process of the Census Bureau's household surveys. These controls are treated as if they were without error – in statistical terms they are treated as though they are unbiased and have zero variance. Survey estimates that are directly controlled to these population estimates are thus also assumed to have zero bias and variance. As noted above some of the age/race/sex detail is derived for each component of change via demographic assumptions or modeling. Thus, even if each administrative records source that is used completely covered the relevant population and reporting was perfect, there would be variability and possibly bias introduced into the detailed estimates. As the level of geography for which estimates are produced moves from the nation to the states and then to counties, the amount of uncertainty and potential bias increases as additional modeling and/or assumptions are required. How can we properly take account of these sources of bias and variation and attach meaningful quantitative measures to them or to the total variation?

The Program of Integrated Estimates (PIE) is a series of projects at the Census Bureau for improving the accuracy and reliability of the population estimates through incorporation of data from the ACS. The ACS will be a single nationwide data source that can provide county and sub-county data to aid in estimating patterns of international and internal migration, changes in housing and racial characteristics, and fertility differentials by race (in addition to providing data for deriving many other types of estimates for many different purposes).

An example of how ACS data can be used to adjust for bias in the use of administrative records is provided by the estimation of internal migration. As noted above, state level migration estimates are currently based on matching tax returns from consecutive years by social security number.

Because some people are not required to file tax returns (generally due to having low income), persons filing returns or claimed as dependents cover only about 80 percent of the population. In migration models the consecutive year filers are assumed to have patterns representative of the entire population. ACS will collect single year migration information directly from its respondents regardless of income, which can be used to adjust the tax-based state migration estimates for differential migration rates between filers and non-filers of tax returns.

With the introduction of the PIE and its incorporation of ACS data into the population estimates, the issue becomes how to derive measures of uncertainty and bias for the combined estimates? If we were to combine data from two surveys, we would generally attempt do so in such a way that we optimized some function of the bias and variance. How do we extend such ideas to the combination of estimates from administrative records with estimates obtained from models and survey data?

4. Demographic Population Projections and Statistical Forecasting

Box and Jenkins (1970, p. 2) state the objectives of statistical forecasting as follows:

> Our objective is to obtain a forecast function which is such that the mean square of the deviations $z_{t+l} - \hat{z}_t(l)$ between the actual $[z_{t+l}]$ and forecasted values $[\hat{z}_t(l)]$ is as small as possible *for each lead time l.*
>
> In addition to calculating the best forecasts, it is also necessary to specify their accuracy, so that, for example, the risks associated with decisions based upon the forecasts may be calculated. The accuracy of the forecasts may be expressed by calculating *probability limits* on either side of each forecast. [italics in original text]

It is reasonable to say that these are the general objectives in statistical forecasting, not just the objectives of Box and Jenkins. Of course, practical limitations mean that these objectives cannot be achieved exactly. In the real world the true model is never known, and so the point forecasts produced provide at best approximations to the optimal forecasts. The associated probability limits will also be approximate. For example, it is generally impossible to account for all sources of error in measures of forecast accuracy – error due to using the wrong model is one source likely to be missing. Nevertheless, statisticians generally accept the above two objectives as their goal.

It is interesting to compare the statistical perspective reflected in the objectives stated above with the perspective of demographers engaged in doing population projections. To gain some insight about the latter, we refer to the conference volume edited by Keilman and Cruijsen (1992). This volume reports on results of a survey of statistical agencies in 30 industrialized countries about their population projections practices. The survey, which was conducted in 1988 by the Netherlands Central Bureau of Statistics (now Statistics Netherlands), the U.S. Census Bureau, and the Netherlands Interdisciplinary Demographic Institute, asked agencies about methods they used to produce projections of births, deaths, and migration. The survey also asked agencies whether they produced alternative projections (variants) and how many, and other questions about the nature of their projections effort and the results provided to users. Though the survey is now 15 years old it probably reveals something about the perspective of demographers in doing population projections that still holds true (and it is 18 years more recent than the Box and Jenkins (1970) reference cited above).

On the surface it may seem that demographers and statisticians share a similar perspective in forecasting. A closer look at actual practice reveals, however, that demographers have been somewhat ambivalent, about the nature of their projections. This is reflected in the following quotations taken from the volume of Cruijsen and Keilman (1992).

> For years, many producers of population projections have argued that they need not be concerned with accuracy. After all, these are projections, not forecasts. Consequently, the results do not claim to foretell events but merely represent the outcome of a mathematical model based on certain assumptions. Such projections represent given scenarios, not predictions … No matter what we call the results of our projection models, we must accept the responsibility that they will be used as forecasts. (Long 1992, pp. 129-130).
>
> Most national agencies claim to produce projections, that is, calculations which show what would happen *if* certain assumptions regarding fertility, mortality and external migration were borne out. By definition, a projection is conditional and it must be correct unless arithmetical errors have been made. But projections are mostly used as forecasts, showing *the most likely* future population trends. Moreover, demographers don't choose *un*likely future fertility,

> mortality or migration trends – unless they explicitly state so (Cruijsen and Keilman 1992, pp. 3-4).

DeBeer (1992, p28) makes similar observations.

Regarding the specification of forecast accuracy, Long (1992, p. 131) noted that, "Perhaps the most common method of facing uncertainty is to produce more than one series of population projections." He also reported (p. 132) that 23 of the 30 countries in the survey reported publishing multiple projections variants. The nature of projection variants is not the same, however, as that of the probability limits suggested above by Box and Jenkins. For example, variants are not generally intended to provide probability statements, although perhaps to many users, they serve the same purpose, that is they say, "we think that the likely outcome is within these bounds …"

Long (1992, p. 131) notes, "These multiple series represent alternative scenarios. Each scenario is based on a different set of explicit assumptions . . ." Also, practices across different countries suggest different magnitudes of uncertainty conveyed by their projection variants. Long (1992, Table 8.1) reports from the survey that the percentage range of fertility projection variants for the year 2000 ranged from 8.4 percent for Cyprus to 45.0 percent for France, with considerable variation in between these extremes among the other countries reporting. While one would expect some country-to-country differences in uncertainty about the future course of fertility, these large differences suggest that the various countries were trying to convey different information about the uncertainty of their projections.

In fact, Long (1992, p. 131) notes two purposes of providing projection variants that are rather different from the purpose behind providing probability limits. One purpose is to let users choose a projection series that most closely reflects their own judgment about the most likely future course of the demographic components. The other purpose is to let users choose a projection series most appropriate for their particular "loss function" regarding the consequences of errors in the projections, as when over-prediction and under-prediction have different consequences for a particular use of the projections. For example, variants are, as discussed below, very much in the spirit of sensitivity analyses. They are not necessarily intended to provide probability statements, although perhaps to many users they serve that purpose. That is, they are interpreted as saying that, "the outcome is likely to be within these bounds." In fact, Keyfitz (1981) and Stoto (1983) did analyses to attach probabilistic measures of uncertainty to official population projections. They analyzed errors in historical U. N. and U.S. Census Bureau population forecasts and used their results in two ways. First, they derived measures of the likely magnitude of the projection errors, and used these to construct confidence intervals for the projected U. S. population in the year 2000. Second, they compared their measures of error to the range defined by U. S. high and low projection variants, which led both to conclude that these variants could be interpreted as defining confidence intervals with probability content of roughly two-thirds.

It should be noted that statisticians also sometimes do sensitivity analyses to reflect uncertainty, particularly in situations where random error is difficult to quantify or the random error that can be quantified is only a small part of the overall error. Of course, demographers often face such situations. Alternative projections seem to be more in the spirit of sensitivity analyses than probability limits.

It should be noted that statisticians also sometimes do sensitivity analyses to reflect uncertainty, particularly in situations where random error is difficult to quantify or the random error that can be quantified is only a small part of the overall error. In these situations, empirical measures that indicate the expected level of uncertainty in forecasting the observed time series may not be terribly useful. Sensitivity analyses that show alternative forecasts under plausible alternative assumptions designed to reflect, to some extent, possible measurement errors, may be more useful. Such situations may come up in population projections, particularly with respect to migration projections.

Though time series models have not yet made much headway in population projections practice, there has been considerable research on the subject. Bell (1997) reviews some relevant papers. Most of these papers focus on use of time series models in forecasting a particular component of demographic change, usually fertility or mortality. Alho and Spencer (1991,1997) have worked on bringing together stochastic forecasts of all the demographic components to produce statistical measures of uncertainty for the forecasts of population. Bell (1992,1997) notes that the high dimensionality of the problem of forecasting age-specific fertility or mortality rates complicates the use of stochastic models. Another difficult problem involves poor data quality, such as errors in migration data mentioned above. Statistical techniques, such as time series models, can do little to account for this source of error.

One other contrast between demographic population projections and statistical forecasting is worth mentioning. National population projections typically involve a relatively long forecast horizon of 50 years or more, though time series data for developing the projections may be relatively short. If one has less than 50 years of data then there is fundamentally no empirical basis for making statements about the population 50 years or more into the future. To state a hypothetical situation, if such data were to contain a cyclical component with a period longer than 50 years this should ideally be accounted for in the projections, yet less than 50 years of historical data will not show even one complete cycle. Clearly, relative to short-term projections, such long-term projections must rest more heavily on assumptions embodied in the model or forecast procedures used, and not so heavily on data.

The focus on the long-term in population projections may strike statisticians as a bit odd given the lack of historical data on which to base the projections. Indeed, most statistical projections tend to increase in uncertainty as they are projected into the long run, eventually becoming uninformative. This is not always the case with demographic projections. Often there are strong assumptions pertaining to the ultimate fertility or mortality rates, with only the speed of the transition to those rates being at issue. The result can be increased uncertainty for the near term, but convergence of the trends in the long run.

For another discussion that gives statistical perspectives on demographic population forecasting, see Daponte, et al. (1997, Section 2).

5. Integrating Demographic Knowledge Into a Statistical Analysis–Is Bayes the Answer?

One way that demographic knowledge may be incorporated into a statistical analysis is through imposing constraints on fundamental quantities. For example, fertility rates and mortality rates must be positive. Constraints seem a rather mild form of knowledge, but if there is sufficient variability in the data they can have important impacts on results, often more so on confidence interval limits than on point estimates. Constraints are often imposed by data transformation, e.g., by taking logarithms of quantities that must be positive. Sometimes demographic knowledge may suggest setting limits that do not correspond to a known natural constraint. For example, Thompson et al. (1989) noted that in the projections reported in U.S. Census Bureau (1988), log (TFR – 1) was modeled to prevent point forecasts and, more importantly, lower forecast limits, of the total fertility rate (TFR) from falling below 1, reflecting a demographic judgment that such low values would not be realized in the U.S. over the course of the forecast horizon. Demographic knowledge may suggest both lower and upper limits on quantities such as TFR. Thompson (1989) and Alho (1990) suggested use of a scaled logistic transformation on the corresponding interval to enforce such interval constraints.

Beyond simple imposition of constraints on estimates, the Bayesian approach provides a formal mechanism for combining subject matter knowledge (expressed in the prior) with information from data. This approach seems to have been little used, however, to integrate demographic knowledge with survey data. One possible reason for this could be the general aversion of statistical agencies to explicit models in general and, thus, to Bayesian treatment of models in particular. Another reason may be that demographic knowledge tends to be expressed about fundamental demographic quantities (e.g., fertility and mortality rates, sex ratios), whereas the standard Bayesian approach wants prior information about model parameters. The fundamental demographic quantities tend to be output variables – things the model would try to predict – not model parameters. The Bayesian approach can, in principle, accommodate this type of prior knowledge, but it is not the standard approach.

We need to clarify here what we have in mind in terms of using a Bayesian approach to combine demographic knowledge with statistical estimates. We have in mind situations where the prior obtained from the demographic knowledge and the information obtained from the data would both have an appreciable impact on the results. Note that this excludes analyses based on non-informative priors, since such priors, by definition, have little or no impact on the results. In such situations the Bayesian approach may still have advantages in terms of how it uses the data, but it effectively has no prior knowledge to combine with the data. We are also excluding analyses based on no data that must rely entirely on prior knowledge. Probabilistic simulation analyses that are purely subjective may be Bayesian, and must reflect prior knowledge, but do not illustrate combination of prior knowledge with statistical data.

A couple of papers have moved in the direction of using a formal Bayesian approach to combine demographic knowledge with statistical estimates. Due to limitations of the prior knowledge or the data, however, they have not quite achieved the degree of integration that we have in mind. One such paper is Elliott and Little (2000). As noted in Section 2, they applied a Bayesian approach to the models of Bell (1993) for combining census coverage survey estimates with demographic analysis sex ratios. The priors used by Elliott and Little were largely noninformative, though, except for the assumption that the national level sex ratios from demographic analysis were truth. Thus, as in Bell (1993), the demographic knowledge involved (sex ratios) was taken as certain and imposed a constraint on the model, rather than being represented as uncertain via a (non-degenerate) prior distribution. Elliott and Little did mention the possibility of extending their approach to recognize uncertainty about the true sex ratios.

Daponte et al. (1997) took an explicitly Bayesian approach to projecting the Iraqi Kurdish population. The goal was to incorporate the "demographer's uncertainty about the past and future characteristics of the population in the form of elicited prior distributions." Unfortunately, the data available were very limited, with most of the data not referring directly to the Iraqi Kurdish population. The data were mostly used to determine prior means around which were placed normal distributions to subjectively reflect the uncertainty about the various demographic components. Having used the limited data to develop the "priors," these also became the posterior distributions since there was no additional data to be combined with the priors. Stochastic population projections were then obtained by simulation.

The paper of Robinson et al. (1993) provides a somewhat related analysis to that of Daponte et al. (1997). Though not labeling their analysis as Bayesian, Robinson et al. subjectively specified probability distributions reflecting uncertainty about demographic components of population change and then produced subjective uncertainty intervals for demographic analysis estimates by simulation. As with Daponte et al., the available data were used in developing the distributions of the demographic components.

Lacking additional examples to illustrate the possibilities, we now consider what is generally likely to happen if demographic knowledge expressed through a prior is combined with data (survey estimates) via a Bayesian approach. First, if the prior is consistent with the data then the two reinforce each other with little effect on point estimates but reduction of variances (relative to not using the prior). In this case presumably both the statistician and the demographer would be satisfied. The remaining scenarios consider what may happen if the demographic prior is inconsistent with the survey estimates.

- If both the demographic knowledge and the data seem relatively weak then one is left with not much information. The statistician may be satisfied if the resulting estimates have high variances, reflecting the substantial uncertainty that exists. The demographer may be satisfied if alternative estimates (variants) reflecting different demographic assumptions can be provided to illustrate the sensitivity of the estimates to the assumptions. For example, this may be the situation with respect to estimation of the undocumented population.

- If the demographic knowledge seems solid and the survey estimates being used seem implausible on their own (suggesting errors in the data), then the effects of the prior may be relatively easy for the statistician to accept. There may be some concern and disagreement, though, about how much effect the prior (and, conversely, the data), should have on the results. This may be the situation for estimation of the population for adult Black males.

- If the demographic knowledge seems questionable and the survey estimates seem fairly reliable then the statistician will be reluctant to let the demographic prior have much effect on the results. This may lead to abandoning the Bayesian approach or at least to using a non-informative prior. Of course, the demographer may disagree about the value of the demographic knowledge. This may be the situation for estimation of the Asian population.

- If the demographic knowledge seems solid (at least in the mind of the demographer) and the survey estimates seem reliable (at least in the mind of the statistician), then conflicts between the prior and the survey estimates will seem harder to accept, and disagreements about the use of the prior knowledge may result. This may be true even if some or all of the difference between the prior and the survey estimates could be attributed to sampling variation. In such a case will the statistician be convinced that use of the prior knowledge is not producing bias in the results? Will the demographer be convinced that there are not significant non-sampling errors producing bias in the survey estimates? An example might be estimation of the population for children under age five.

Our point here is that while the Bayesian approach provides a formal mechanism for producing an answer that combines prior knowledge with new data (in the form of a posterior distribution), it does not guarantee agreement between the subject matter expert (the demographer) and the statistician.

6. Conclusions and Future Research

Where do we go from here? Demographers need first to develop and document more explicit models so that their analyses are more transparent. Great progress has already been made here, at least compared to the early, pre-computer, days. More fundamentally, demographers need to really identify the variability in their models and the uncertainty in their results. This will be no easy task given both the high dimensionality of the problems and the multiplicity of the data sources. They need to think about what the concepts of bias and variance might mean to their problems or whether other concepts need to be developed, concepts that categorize and analyze uncertainty from other perspectives.

Statisticians need to continue work on methods to combine data of different types and from different sources when such data are of quite different quality and have different error structures. Work on quantifying nonsampling errors and estimating biases needs to continue. Thus, just as demographers should strive to develop objective measures of errors in their estimates (when possible), statisticians should strive to account for both sampling and non-sampling errors in their measures of uncertainty.

Together, statisticians and demographers need to develop an overall framework or paradigm to conduct this discussion. The Bayesian model may be a good one. There are others. Economics and statistics have a combined approach – econometrics – that strives to integrate theory and substantive knowledge on the one hand with empirical methods on the other hand. This brings us back to the beginning. Demography started as the most quantitative of the sciences. It started hand-in-hand with statistics. The two seem to have gone their separate ways, but obviously need to come back together.

References

Alho, Juha M. (1990), "Stochastic Methods in Population Forecasting," *International Journal of Forecasting*, **6**, 521-530.

Alho, Juha M. and Spencer, Bruce D. (1991), "A Population Forecast as a Database: Implementing the Stochastic Propagation of Error", *Journal of Official Statistics*, **7**, 295-310.

Alho, Juha M. and Spencer, Bruce D. (1997), "The Practical Specification of the Expected Error of Population Forecasts", *Journal of Official Statistics*, **13**, 203-225.

Anderson, Margo (ed) (2000), *Encyclopedia of the U. S. Census*, Washington D.C.: CQ Press.

Bell, William R. (1992), "ARIMA and Principal Component Models in Forecasting Age-Specific Fertility," Chapter 10 of *National Population Forecasting in Industrialized Countries*, eds. Nico Keilman and Harri Cruijsen, Amsterdam: Swets and Zeitlinger, 178-200.

Bell, William R. (1993), "Using Information From Demographic Analysis in Post-Enumeration Survey Estimation,"*Journal of the American Statistical Association*, **88**, pp. 1106-1118.

Bell, William R. (1997), "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates," *Journal of Official Statistics*, **13**, 279-303.

Box, G.E.P. and Jenkins, G. M. (1970), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.

Coale, Ansley and F. Stephan (1962), "The Case of the Indians and the Teen-Age Widows," *Journal of the American Statistical Association*, **57**, 338-347.

Coale, Ansley and Melvin Zelnick (1963) *New Estimates of Fertility and Population in the United States*, Princeton University Press, 1963.

Cruijsen, Harri and Keilman, Nico (1992), "A Comparative Analysis of the Forecasting Process," Chapter 1 of *National Population Forecasting in Industrialized Countries*, eds. Nico Keilman and Harri Cruijsen, Amsterdam: Swets and Zeitlinger, 3-25.

Daponte, Beth Osborne, Kadane, Joseph B, and Wolfson, Lara J. (1997), "Bayesian Demography: Projecting the Iraqi Kurdish Population, 1977-1990," *Journal of the*

American Statistical Association, **92**, 1256-1267.

DeBeer, Joop (1992), "General Time-Series Models for Forecasting Fertility," Chapter 9 of *National Population Forecasting in Industrialized Countries*, eds. Nico Keilman and Harri Cruijsen, Amsterdam: Swets and Zeitlinger, 147-176.

Elliott, Michael R. and Little, Roderick J.A. (2000), "A Bayesian Approach to Combining Information from a Census, a Coverage Measurement Survey and Demographic Analysis," *Journal of the American Statistical Association*, **95**, 351-362.

Fay, Robert III, Passel, Jeffery and Robinson, J. Gregory (1988) *The 1980 Census, PHC80-E4*, U.S. Census Bureau, Washington, DC.

Hollman, Frederick W., Lisa B. Kuzmeskus, R. Colby Perkins, and Elizabeth Weber (1998), "United States Population Estimates by Age, Sex, Race and Hispanic Origin: 1990 to 1997," *PPL-91*, U.S. Census Bureau, Washington, DC

Keilman, Nico and Cruijsen, Harri eds. (1992), *National Population Forecasting in Industrialized Countries*, Amsterdam: Swets and Zeitlinger.

Keyfitz, Nathan (1981), "The Limits of Population Forecasting," *Population and Development Review, **7**, 579-593.*

Long, John F. (1992), "Accuracy, Monitoring, and Evaluation of National Population Projections," Chapter 8 of *National Population Forecasting in Industrialized Countries*, eds. Nico Keilman and Harri Cruijsen, Amsterdam: Swets and Zeitlinger, 129-145.

Pearson, E. S. and J. W. Tukey (1965), "Approximate means and standard deviations based on distances between percentage points of frequency curves," *Biometrika*, **52**, 533-546.

Robinson, J. Gregory, Ahmed, Bashir, Das Gupta, Prithwis, and Woodrow, Karen A. (1993), "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis (with discussion)," *Journal of the American Statistical Association*, **88**, 1061-1079.

Robinson, J. G. (2001), "ESCAP II: Demographic Analysis Results," Executive Steering Committee for A.C.E. Policy II, Report 1, U.S. Census Bureau, Washington, DC.

Robinson, J. G. (2001), "Accuracy and Coverage Evaluation Survey: Demographic Analysis Results," Executive Steering Committee for A.C.E. Policy, Report B-4, U.S. Census Bureau, Washington, DC.

Robinson, J Gregory, 2002 "Coverage of Population in Census 2000: Results from Demographic Analysis," paper presented to the Population Association of America.

Siegel, Jacob (1974) *Census, PHC (E)-4*, "Estimates of Coverage of Population by Sex, Race and Age: Demographic Analysis," 15-18, U.S. Census Bureau, Washington, DC..

Siegel, Jacob, Passel Jeffrey and Robinson, J. Gregory (1977), *Current Population Reports, Series P-23, No. 65*, "Developmental Estimates of the Coverage of the Population of States in the 1970 Census: Demographic Analysis," U.S. Government Printing Office, Washington, D.C.

Sink, Larry (1999), "Producing County-Level Characteristics Estimates by the Cohort-Component Model," paper presented at the Southern Demographic Association meetings.

Stoto, M.A. (1983), "The Accuracy of Population Projections," *Journal of the American Statistical Association*, **78**, 13-20.

Thompson, Patrick A. (1989), "A Transformation Useful for Bounding a Forecast," *Statistics and Probability Letters*, **8**, 469-475.

Thompson, Patrick A., Bell, William R., Long, John. F., and Miller, Robert B. (1989), "Multivariate Time Series Projections of Parameterized Age-Specific Fertility Rates," *Journal of the American Statistical Association*, **84**, 689-699.

U.S. Census Bureau (1977), Estimates of Coverage of Population by Sex, Race and Age: Demographic Analysis, *1970 Census Report PHC (E)-4, pp. 15-18,* Washington, DC.

U.S. Census Bureau (1988), "Projections of the Population of the United States, by Age, Sex, and Race: 1988-2080," *Current Population Report 952 (Ser. P-25)*, Washington, DC, U.S. Government Printing Office.

U.S. Census Bureau, (2001) "General Methodology – U.S. Population Estimates by Age, Sex, Race, and Hispanic Origin: 2001," http://eire.census.gov/popest/topics/methodology/asro.php.

U.S. Census Bureau, "State and County Total Population Estimates," http://eire.census.gov/popest/topics/methodology/states.php.

Wolter, Kirk (1986) "Some Coverage Error Models for Census Data" *Journal of the American Statistical Association*, **81**, 338-346.

Table 1. Comparison of Initial and Revised Estimates of Percent Net Undercount of the US Resident Population Based on Demographic Analysis, by Sex and Race: 1970, 1980, 1990 and 2000

Census	Row	Publication Date	Total	Male	Female	Black	Nonblack
	1	Feb 1974	2.55	3.30	1.80	7.70	1.80
1970	2	Sept. 1985	2.92	3.70	2.16	8.02	2.24
	3	Apr. 1991	2.71	3.45	1.99	6.48	2.21
	4	Feb. 1982	-0.40	0.50	-1.20	4.80	-1.10
1980	5	Sept. 1985	1.38	2.37	0.42	5.89	0.75
	6	Apr. 1991	1.22	2.20	0.28	4.50	0.77
1990	7	Sept. 1993	1.85	2.79	0.94	5.68	1.29
	8	Oct. 2001	1.65	2.39	0.93	5.52	1.08
	9	Mar. 2001	-0.65	-0.13	-1.16	2.8	-1.19
2000	10	Mar. 2001	0.32	0.91	-0.25	3.51	-0.17
	11	Oct. 2001	0.12	0.86	-0.60	2.78	-0.29

Row. 1 - 1970 Census, PHC (E)-4 (Feb. 1974), Table 3
Row 2, 5 - 1980 Census, PHC80-E4 (Feb. 1988), Table 3.2
Row 3, 6, 7 - JASA, Vol. 88, No. 423 (Sept. 1993), Table 2
Row 4 - Current Population Reports, Series P-23, No. 115 (Feb. 1982), Table 1
Row 8, 11 - ESCAP II, Report No. 1 (Oct. 2001), Table 4 and 6
Row 9, 10 - ESCAP I, Report B-4 (March 2001), Table 4 and 6

Table 2. Comparison of Initial and Revised Estimates of Percent Net Differential Undercount Based on Demographic Analysis, by Sex and Race: 1970, 1980, 1990 and 2000

Census	Publication Date		Male minus Female	Black minus Nonblack
	Feb.	1974	1.50	5.90
1970	Sept.	1985	1.54	5.78
	Apr.	1991	1.46	4.27
	Feb.	1982	1.70	5.90
1980	Sept.	1985	1.95	5.14
	Apr.	1991	1.92	3.73
1990	Sept.	1993	1.85	4.39
	Oct.	2001	1.46	4.44
	Mar.	2001	1.03	3.99
2000	Mar.	2001	1.16	3.68
	Oct.	2001	1.46	3.07