MEASURING UNCERTAINTY IN POPULATION FORECASTS: A NEW APPROACH

David A. Swanson (Correspondence author)
Department of Sociology and
Center for Sustainable Suburban Development
University of California Riverside
Riverside, California 92521 USA
E-mail: David.swanson@ucr.edu

Jeff Tayman
Department of Economics
University of California San Diego
9500 Gilman Drive
La Jolla, California 92093-0508 USA
Email: jtayman@ucsd.edu

OUTLINE

1. Overview
2. Cohort Change Ratios (CCRs)
3. Regression-Estimated CCRs
4. Measuring Uncertainty
5. Data
6. Results
7. Discussion
8. Works Cited in this Presentation

OVERVIEW

Two basic approaches have been used to assess population forecast uncertainty: (1) a range of projections based on alternative scenarios; and (2) statistical forecast intervals. In terms of the latter, there are two complementary approaches: (1) model-based intervals; and (2) empirically-based intervals.

We evaluate a model-based approach in this paper, but enhance it by using historical data, a feature found in the empirically-based approach.

OVERVIEW

We describe and test a regression-based approach for developing 66% forecast intervals for age-group forecasts made using the Hamilton-Perry Method. To evaluate this method, we use 16 age groups (0-4, 5-9,…, 70-74, 75+) taken from a sample of four states (one from each census region in the United States) and nine ex post facto tests, one for each census from 1930 to 2010. This yields 576 observations for which we can see if the forecast interval for a given age group in a given census year contains the census count for the same age group.

OVERVIEW

The four states and the nine target years provide a wide range of characteristics in regard to population size, growth, and age-composition, factors that affect forecast accuracy. The tests reveal that the 66% intervals contain the census age-groups in 397 of the 576 observations (69 percent). We discuss the results, including a summary by age group, and make some observations regarding the limitations of our study. We conclude that the results are encouraging, however, and offer suggestions for further work.

COHORT CHANGE RATIOS

What are Cohort Change Ratios (CCRs)? They have a long history of use in demography. Under the rubric of "Census Survival Ratios," they have been used to estimate adult mortality and under the rubric of the "Hamilton-Perry" method, to make population projections.

COHORT CHANGE RATIOS

The Hamilton-Perry Method is a variant of the cohort component method that has far less intensive input data requirements. Instead of mortality, fertility, migration, and total population data, which are required by the full-blown cohort-component method, the Hamilton-Perry method requires data only from the two most recent censuses.

COHORT CHANGE RATIOS

The Hamilton-Perry method projects a population by age (and sex) from time (t) to time (t+k) using CCRs computed from the two most recent censuses. It consists of two steps. The first uses existing data to develop CCRs and the second applies the CCRs to the cohorts of the launch year population to move them into the future. The second step can be repeated infinitely, with the projected population serving as the launch population for the next projection cycle.

COHORT CHANGE RATIOS

The formula for the first step, the development of a CCR is:

$$ {}_nCCR_x = {}_nP_{x,t} / {}_nP_{x-k,t-k} $$

where

${}_nP_{x,t}$ is the population aged x to x+n at the most recent census (t),

${}_nP_{x-k,t-k}$ is the population aged x-k to x-k+n at the 2nd most recent census (t-k),

k is the number of years between the most recent census at time t and the one preceding it at time t-k.

COHORT CHANGE RATIOS

The basic formula for the second step, projecting age cohorts is:

$$ {}_nP_{x+k,t+k} = ({}_nCCR_x) \times ({}_nP_{x,t}) $$

where

${}_nP_{x+k,t+k}$ is the population aged x+k to x+k+n at time (t+k), and

${}_nP_{x,t}$ is the population aged x to x+n at the most recent census (t).

COHORT CHANGE RATIOS

Given the nature of the CCRs, 10-14 is the youngest age group for which projections can be made if there are 10 years between censuses. To project the populations aged 0-4 and 5-9, one can use the Child Woman Ratio (CWR), or more generally a “Child Adult Ratio” (CAR). It does not require any data beyond the decennial census. For projecting the population aged 0-4, CAR is defined as the population aged 0-4 divided by the population aged 15-44. For projecting the population aged 5-9, CAR is defined as the population aged 5-9 divided by the population aged 20-49.*

*****There are other “adult” age groups that could be used to define CAR.

COHORT CHANGE RATIOS

Another way to obtain “CCRs” for the two youngest age groups is to take their ratios at two points in time and apply that ratio to the launch year age group (*t*). In the first step, the ratios are :

Population 0-4: $_5R_{0,t} = {}_5P_{0,t} / {}_5P_{0,t-k}$

Population 5-9: $_5R_{5,t} = {}_5P_{5,t} / {}_5P_{5,t-k}$.

In the second step, the projected population at *t*+*k* is found by:

Population 0-4: $_5P_{0,t+k} = {}_5P_{0,t} \times {}_5R_{0,t}$

Population 5-9: $_5P_{5,t+k} = {}_5P_{5,t} \times {}_5R_{5,t}$.

COHORT CHANGE RATIOS

We use the later method since it is better suited for the regression-based method for creating intervals around forecasts for the two youngest age groups discussed later in the paper.

One reason that it is better suited with the regression-based method is that the CAR values are substantially different than the CCRs, whereas the ratios are not. This means that the CAR values are potential outliers that could serve as influential observations that deleteriously affect model construction.

COHORT CHANGE RATIOS

Projections of the oldest open-ended age group differ slightly from the CCR projections for the age groups beyond age 10 up to the oldest open-ended age group. If, for example, the final closed age group is 70-74, with 75+ as the terminal open-ended age group, then calculations for the ${}_{\infty}CCR_{75,t}$ require the summation of the three oldest age groups to get the population age 75+ at time t and the summation of the age groups that will yield P65+ at time t-k:

$${}_{\infty}CCR_{75,t} = {}_{\infty}P_{75+,t} / {}_{\infty}P_{65+,t-k}.$$

The formula for projecting the population 75+ for the year t+k is: ${}_{\infty}P_{75+,t+k} = ({}_{\infty}CCR_{75+,t}) \times ({}_{\infty}P_{65+,t})$.

COHORT CHANGE RATIOS

Since the population data are five-year age groups with a final open-ended age group of 75+, the conventions described above are used in terms of the CCRs and the projections of the youngest two age groups (0-4 and 5-9) and the terminal open-ended age group (75+). Important to the subsequent discussion are the CCRs developed from the 1990 and 2000 census data by age.

COHORT CHANGE RATIOS

Table A2.1 (taken from our full paper) provides an example of the Hamilton-Perry Method for Minnesota (one of our four sample states). We won't discuss its details at this time, but we want to show an example of the and the generation of the CCRs from 1990 and 2000 census data by age.

Table A2.1 Ratios, 1980-1990 and 1990-2000 and Projected Population 2010, Minnesota

	Population			Ratios[a]			
					1990-2000		2010
Age	1980	1990	2000	1980-1990	Observed	Estimated[b]	Population
0 to 4	307,249	336,800	329,594	1.09618	0.97860	1.11501	367,501
5 to 9	296,295	345,840	355,894	1.16722	1.02907	1.17641	418,677
10 to 14	333,378	313,297	374,995	1.01968	1.11341	1.04890	345,711
15 to 19	399,818	297,609	374,362	1.00443	1.08247	1.03572	368,607
20 to 24	393,566	316,046	322,483	0.94801	1.02932	0.98696	370,105
25 to 29	363,435	381,759	319,826	0.95483	1.07465	0.99286	371,689
30 to 34	313,104	397,984	353,312	1.01123	1.11791	1.04160	335,898
35 to 39	246,356	361,274	412,490	0.99405	1.08050	1.02675	328,381
40 to 44	202,860	304,810	411,692	0.97351	1.03444	1.00900	356,492
45 to 49	187,051	237,050	364,247	0.96223	1.00823	0.99925	412,181
50 to 54	193,199	191,410	301,449	0.94356	0.98897	0.98312	404,743
55 to 59	189,457	173,066	226,857	0.92523	0.95700	0.96727	352,325
60 to 64	170,638	171,220	178,012	0.88624	0.93000	0.93358	281,427
65 to 69	149,114	160,036	153,169	0.84471	0.88503	0.89769	203,647
70 to 74	121,034	134,486	142,656	0.78814	0.83317	0.84880	151,097
75+	209,416	252,412	298,441	0.52634	0.54566	0.62254	369,954
Total	4,075,970	4,375,099	4,919,479				5,438,435

[a] Ages 0-4 = $P_{0\text{-}4,t}$ / $P_{0\text{-}4,t\text{-}10}$.
Ages 5-9 = $P_{5\text{-}9,t}$ / $P_{5\text{-}9,t\text{-}10}$.
Ages 10-74 = $P_{x+10,t}$ / $P_{x,t\text{-}10}$.
Ages 75+ = $P_{75+,t}$ / $P_{65+,t\text{-}10}$.

[b] Based on the regression equation, $0.1676667 + (0.8644256 \times \text{Ratios}_{1980\text{-}1990})$

[c] Ages 0-4 = Est.1990-2000 $\text{Ratio}_{0\text{-}4} \times P_{0\text{-}4,2000}$.
Ages 5-9 = Est.1990-2000 $\text{Ratio}_{5\text{-}9} \times P_{5\text{-}9,2000}$.
Ages 10-14 = Est.1990-2000 $\text{CCR}_{x} \times P_{x\text{-}10,2000}$.
Ages 75+ = Est.1990-2000 $\text{CCR}_{75+} \times P_{65+,2000}$.

REGRESSION-ESTIMATED CCRs

The Hamilton-Perry Method is deterministic, which is not surprising given its consistency with the fundamental demographic accounting equation. However, we also know that population forecasting is subject to uncertainty since we do not precisely know the future components making up the fundamental equation. So, the question is how to introduce an element of statistical uncertainty into a method that is inherently deterministic. One answer is found by employing regression techniques to forecast CCRs and their intervals.

REGRESSION-ESTIMATED CCRs

Recall that ${}_nCCR_{x,t} = {}_nP_{x,t} / {}_nP_{x-k,t-k}$.

From this, we can define the CCR for the preceding census period as ${}_nCCR_{x,t-k} = {}_nP_{x,t-k} / {}_nP_{x-k,t-2k}$.

We then construct a regression model with ${}_nCCR_{x,t}$ as the dependent variable and ${}_nCCR_{x,t-k}$ as the independent variable.

For age groups 0-4, 5-9, and the terminal open-ended age group that the dependent and independent observations follow the equations provided earlier.

REGRESSION-ESTIMATED CCRs

Given this adjustment, we estimate the CCRs at time (t) by:

$$ {}_nECCR_{x,t} = a + b \times {}_nCCR_{x,t-k}. $$

We then multiply the regression-estimated CCR and the corresponding population by age at time (t) to forecast the CCR at time (t+k):

$$ {}_nCCR_{x,t+k} = {}_nECCR_{x,t} \times {}_nP_{x,t}. $$

MEASURING UNCERTAINTY

Utilizing the regression measure of statistical uncertainty (the standard error of estimate) for the model along with the sample size and other characteristics of the data, we generate forecast intervals around ${}_nCCR_{x,t+k}$. based on equation 4.2 found in Hyndman and Athanasopoulos, Chapter 4 (2012). These intervals can be translated directly to the forecasted population numbers for each age group (Espenshade and Tayman 1982; Swanson and Beck 1994).

MEASURING UNCERTAINTY

When the prediction from a regression equation is derived from an observed data value, we call the resulting value of a “fitted value.” This is not a forecast as the actual value of a predictor variable is used in the calculation. When values of the predictor variable are not part of the data used to estimate the model, the resulting prediction is a forecast.

Assuming that the regression errors are normally distributed, an approximate 95% **forecast interval** (also called a prediction interval) associated with this forecast is given by Hyndman and Athanasopoulos, Chapter 4,2012) as.

$$\hat{y} \pm 1.96 s_\varepsilon \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{(N-1)s_x^2}},$$

DATA

To empirically examine the regression-based method for developing intervals around population forecasts by age generated from the Hamilton-Perry Method, we selected a sample made up of one state from each of the four census regions in the United States. The states selected are Georgia (the South Region), Minnesota (the Midwest Region), New Jersey (The Northeast Region) and Washington (The West Region).

DATA

We assembled census data for these four states for each census year from 1900 to 2010. The data provide nine time points at which the forecast intervals can be evaluated, 1930, 1940, 1950, 1960, 1970, 1980, 1990, 2000, and 2010. This sample provides a wide range of demographic characteristics in terms of variation in population size, age-composition, and rates of change.

DATA

Table 1 provides an overview of this range by displaying the population of each of the four states in 1900 and in 2010 and decennial rates of population change from 1900 to 2010. Although we do not show a summary of the changes in age composition by state and census year, they are extensive as seen in Appendix 1 of the full paper, which provide the age data by state and census year.

Table 1. Total Population 1900 and 2010 and
Annual Rate of Change by Decade, Sample States

Census Year	Georgia	Minnesota	New Jersey	Washington
1900[a]	2,209,974	1,747,292	1,879,890	511,844
1900-1910	1.64%	1.70%	2.99%	7.97%
1910-1920	1.05%	1.41%	2.19%	1.75%
1920-1930	0.05%	0.72%	2.47%	1.44%
1930-1940	0.72%	0.86%	0.30%	1.06%
1940-1950	0.98%	0.66%	1.50%	3.14%
1950-1960	1.35%	1.35%	2.27%	1.83%
1960-1970	1.52%	1.08%	1.67%	1.78%
1970-1980	1.74%	0.69%	0.27%	1.92%
1980-1990	1.70%	0.71%	0.48%	1.64%
1990-2000	2.34%	1.17%	0.85%	1.92%
2000-2010	1.68%	0.75%	0.44%	1.32%
2010	9,687,653	5,303,925	8,791,894	6,724,540

[a] The 1900 population totals exclude those for whom age was not reported.

DATA

We constructed CCRs over two successive decennial periods (e.g., 1910-1920/1900-1910) over the entire period, using regression to estimate the CCR in the numerator from the CCR in the denominator.

We then used the regression-based estimate of the CCR of the “current period” (e.g., 1910-1920) to forecast the CCRs to the next period, the “launch year” (e.g., 1920-1930) and developed forecast intervals around these forecasted CCRs, which are then translated into the forecasted age groups for the “target year” (e.g., 1930).

The forecast intervals are then examined to see if they contain the census age groups for the target year.

RESULTS

How well does the regression approach based on the Hamilton Perry method perform in its ability to predict the uncertainty of population forecasts?

One way to address this question is to determine the number of population counts that fall inside the forecast intervals (Tayman, Smith, and Lin 2007). In terms of the forecast interval probability, we selected 0.66 or 66 percent because of prior research indicating that “low” and “high” scenarios constructed for the cohort-component method corresponded empirically to 66% confidence intervals (Stoto 1983) as well as findings by Swanson and Beck (1994).

RESULTS

Table 2 provides a summary of the results for all four states at each of the nine census test points. The table shows the number of times (out of 16) that the 66% forecast interval contained the corresponding census number for a given age group. If the forecast intervals provide a valid measure of uncertainty, they will contain approximately 11 of the 16 observed population counts.

The table also shows percent of the counts falling within the forecast intervals for all target years for each state (144 intervals), the percent falling within all states for each target year (64 intervals), and the single percent falling within all states for all target years (576 intervals).

Table 2. Number of Population Counts Falling within the 66% Forecast Intervals by State and Target Year

Target Year	Georgia	Minnesota	New Jersey	Washington	Total	Percent (N/64)
1930	9	12	8	13	42	67%
1940	3	5	11	12	31	48%
1950	10	14	4	3	31	47%
1960	13	14	14	8	49	86%
1970	6	12	14	13	45	77%
1980	7	12	12	10	41	67%
1990	13	14	14	14	55	83%
2000	8	15	14	15	52	81%
2010	7	15	15	14	51	81%
Total	76	113	106	102	397	
Percent	53%	78%	74%	71%	69%	
	Percent (N/144)	Percent (N/144)	Percent (N/144)	Percent (N/144)	Percent (N/576)	

RESULTS

Table 3 contains a summary of the results by age group across all of the nine census target years and the four states. The table shows the number of times (out of 36) that the 66% forecast interval contained the corresponding census number for a given age group. If the forecast intervals provide a valid measure of uncertainty, they will contain approximately 24 of the 36 observed population counts.

In general, Table shows that forecast intervals capture the population count at least 66 percent of the time for age groups 10-14, 15-19, 20-24 and 40-44 through 75+.

RESULTS

For age groups 0-4 and 5-9, the forecast intervals only encompass the population counts 25 percent of time. For age group 30-34, the count is encompassed 53 percent of the time while for age group 25-29, it is 58 percent of the time. The population counts are captured by the forecast intervals 61 percent of the time for age group 35-39.

Table 3. Number of Population Counts Falling within the 66% Forecast Interval by Age Group

Age	Number	Percent (N/36)
0 to 4	9	25%
5 to 9	9	25%
10 to 14	26	72%
15 to 19	27	75%
20 to 24	24	67%
25 to 29	21	58%
30 to 34	19	53%
35 to 39	22	61%
40 to 44	26	72%
45 to 49	28	78%
50 to 54	30	83%
55 to 59	31	86%
60 to 64	30	83%
65 to 69	31	86%
70 to 74	33	92%
75+	31	86%
Total	397	69%

DISCUSSION

Overall, the 66 percent intervals contain their corresponding census age groups in 397 cases, which represents 69 percent of the 576 total observations. In terms of the nine census target years, the overall results show that in five of them (1960, 1970, 1990, 2000, and 2010) the forecast intervals contain the census age groups substantially more than 66 percent of the time. In two target years (1930 and 1980), the intervals contain the census age groups 67 percent of the time.

DISCUSSION

In the remaining two target years, 1940 and 1950, the intervals contain the census age groups 48 percent and 47 percent of the time, respectively. The 1940 test point encompasses the economic boom experienced in the 1920s and the economic depression during the 1930s and the large scale “baby bust” associated with it. The 1950 point encompasses the depression and baby bust period of the 1930s and the economic recovery stimulated by World War II and the initial part of the large scale “baby boom” from 1946 to 1950.

DISCUSSION

In regard to Table 3 and the summary of results by age group, it should not be surprising that the cohort change method is better able to capture older age groups than the very youngest since births are not part of a cohort change ratio. In addition, migration likely comes into play in that the population in the two youngest age groups (0-4 and 5-9) would be moving with their parents, who are likely to be in age groups 25-29, 30-34, and 35-39, the other age groups for which the forecast intervals encompassed the population counts less than 66 percent of the time.

DISCUSSION

Overall, we find that these effects are consistent with theory regarding migration in that those who tend to move are less socially integrated into communities than those who tend not to move and that as adults age, community social integration tends to increase (Goldscheider 1978). Finally, as shown at the bottom of Table 3, the intervals capture the population count 69 percent of the time (397 out of 576), which matches the summary for Table 2.

DISCUSSION

Although they are not shown here, the average width of the forecast intervals appears to us to be reasonable at the 66 percent level in that they are neither so wide as to be meaningless nor too narrow to be overly-restrictive. This is largely consistent with prior work by Swanson and Beck (1994) on confidence intervals derived from regression-based forecasts. Also consistent with the work by Swanson and Beck (1994), is the fact that the regression-based forecast intervals contain the actual numbers by age in 69 percent of the 576 observations provide further support that 66 percent forecast intervals based on the regression-estimated CCR approach are both useful and feasible. We find these results encouraging.

DISCUSSION

At this point, we suggest caution using this method beyond a ten-year forecast horizon. This is consistent with observations about the use of the Hamilton-Perry method in general (Smith, Tayman, and Swanson 2001; Swanson, Schlottmann, and Schmidt 2010) and as such is not a major limitation. We also suggest that this approach to developing uncertainty measures be used with care when applied to small populations, such as those found at the county and sub-county levels.

DISCUSSION

While our sample provides a wide range of demographic behavior in terms of size, age composition, and population changes, it is a sample of states, which means that greater variability in demographic characteristics found at sub-state levels is not present (Swanson, Schlottmann, and Schmidt 2010). We suggest that further research using this approach examine both longer forecast horizons and smaller populations (i.e., the sub-state populations) and different probability intervals. Another area for further research would be to utilize Keyfitz's (1981) approach using root mean square errors in conjunction with the Hamilton-Perry Method.

DISCUSSION

The fact that the forecast intervals do not contain the population counts at least 66 percent of the time for neither the two youngest age groups (0-4 and 5-9) nor the age groups associated with those most likely to be the parents of these children (25-29, 30-34 and 35-39) should not be surprising: The dynamics of birth and migration are difficult to capture in a full-blown cohort-component method forecast and the Hamilton-Perry Method is a variant of the full- blown method (Smith, Tayman and Swanson 2001; Smith and Tayman 2003). Thus, work on these issues in regard to one of these two methods should be of use to the other.

Works Cited in this Presentation

Espenshade, T. and Tayman, J. (1982). Confidence intervals for postcensal population estimates. *Demography* 19 (2): 191-210.

Goldscheider, C. (1978). *Modernization, migration, and urbanization*. Paris, France: International Union for the Scientific Study of Population.

Hyndman, R.J., and Athanasopoulos, G. (2012) *Forecasting: Principles and Practice* (online at http://otexts.com/fpp/)

Keyfitz, N. (1981). The limits of population forecasting. *Population and Development Review* 7: 579-593.

Works Cited in this Presentation

Smith, S.K., Tayman, J., and Swanson D.A. 2001. *Population projections for state and local areas: Methodology and analysis*. New York, NY: Kluwer Academic/Plenum Press.

Stoto, M. (1983). The accuracy of population projections. *Journal of the American Statistical Association* 78: 13-20.

Swanson, D.A. and Beck, D. (1994). A new short-term county population projection method. *Journal of Economic and Social Measurement* 20: 1-26.

Works Cited in this Presentation

Swanson, D.A., Schlottmann, A., and Schmidt, R. (2010). Forecasting the population of census tracts by age and sex: An example of the Hamilton–Perry method in action. *Population Research and Policy Review* 29: 47-63.

Tayman, J., Smith, S.K., and Lin, J. (2007). Precision, bias, and uncertainty for state population forecasts: An exploratory analysis of time series models. *Population Research and Policy Review* 26 (3): 347-369.