J Pop Research (2011) 28:225–243
DOI 10.1007/s12546-011-9054-5

MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts

David A. Swanson · Jeff Tayman · T. M. Bryan

Published online: 18 March 2011

Abstract Accurately measuring a population and its attributes at past, present, and future points in time has been of great interest to demographers. Within discussions of forecast accuracy, demographers have often been criticized for their inaccurate prognostications of the future. Discussions of methods and data are usually at the centre of these criticisms, along with suggestions for providing an idea of forecast uncertainty. The measures used to evaluate the accuracy of forecasts also have received attention and while accuracy is not the only criterion advocated for evaluating demographic forecasts, it is generally acknowledged to be the most important. In this paper, we continue the discussion of measures of forecast accuracy by concentrating on a rescaled version of a measure that is arguably the one used most often in evaluating cross-sectional, subnational forecasts, Mean Absolute Percent Error (MAPE). The rescaled version, MAPE-R, has not had the benefit of a major empirical test, which is the central focus of this paper. We do this by comparing 10-year population forecasts for U.S. counties to 2000 census counts. We find that the MAPE-R offers a significantly more meaningful representation of average error than MAPE in the presence of substantial outlying errors, and we provide guidelines for its implementation.

Keywords MAPE · MAPE-R · National county test · Forecast accuracy

D. A. Swanson (✉)
Department of Sociology, University of California Riverside, Riverside, CA 92521, USA
e-mail: David.swanson@ucr.edu

J. Tayman
Department of Economics, University of California San Diego, La Jolla, CA 92093, USA

T. M. Bryan
McKibben Demographic Research, PO Box 2921, Rock Hill, SC 29732, USA

Introduction

Any summary measure of error should meet five desirable criteria: measurement validity, reliability, ease of interpretation, clarity of presentation, and support of statistical analysis (National Research Council 1980). The most often used summary measure of forecast error, the MAPE (Mean Absolute Percent Error), meets most of these criteria with one important exception, the lack of measurement validity. Being an arithmetic mean it is affected by extreme values and often overstates the error represented by most of the observations in a population forecast. For the MAPE, extreme values occur only at the high end because it is typically based on a right-skewed distribution of absolute percentage errors bounded on the left by zero and unbounded on the right. In a comprehensive analysis of county-level projections, the MAPE was on average higher by about 30–40% than robust measures of central tendency for most methods and projection horizons (Rayer 2007).

The upward bias of the MAPE is unfortunate because accuracy still remains the most important forecast evaluation criterion (Yokum and Armstrong 1995). Because of the shortcomings of the arithmetic average in an asymmetrical distribution, statisticians suggest alternative measures. The median is one such alternative, but it ignores most of the information and has less desirable statistical properties than an arithmetic mean. Other alternatives include the geometric mean, the symmetrical MAPE (SMAPE), and robust M-estimators. Tayman and Swanson (1999) found that M-estimators more accurately reflected the overall error in a set of forecasts, but they lack the intuitive, interpretative qualities of the MAPE and are unfamiliar to many users and producers of population forecasts. They also found that SMAPE, which is a linear transformation of the MAPE, was not a suitable alternative to the MAPE and suggested the use of non-linear transformations to transform the skewed absolute percentage error distribution into a symmetrical one.

In this paper, we focus on a rescaled version of the MAPE. The rescaled version, MAPE-R, was introduced by Tayman et al. (1999), given a limited empirical test by Swanson et al. (2000), and conceptually and computationally refined by Coleman and Swanson (2007). MAPE-R is based on a power transformation of the error distribution underlying the MAPE. It is designed to address the impact of outlying errors on the MAPE, which can overstate the error represented by 'most' of the observations, while still preserving the valuable statistical properties of an average. However, MAPE-R has not had the benefit of a major empirical test, which is the central focus of this paper, along with providing criteria for its implementation. The initial investigations of MAPE-R included empirical illustrations in the form of case studies for estimates (Swanson et al. 2000) and for forecasts (Tayman et al. 1999). However, the case study data were not intended to provide a comprehensive empirical portrait of MAPE-R, its features and characteristics, which this paper is designed to do.

Measures of forecast accuracy

Swanson and Stephan (2004) define a population forecast as

> … an approximation of the future size of the population for a given area, often including its composition and distribution. A population forecast usually is one of a set of projections selected as the most likely representation of the future. (Swanson and Stephan 2004, p. 770).

Using this definition of a forecast, population forecast error is then the difference between the observed and the forecast population at a designated point in forecast period; that is $E = F - O$. This follows a long-standing tradition of using the *ex-post facto* perspective in examining forecast error, where the error of a forecast is evaluated relative to what was subsequently observed, typically a census-based benchmark (Campbell 2002; Mulder 2002). Forecast errors can be evaluated ignoring the direction of error or accounting for its direction. Measures based on the former evaluate forecast precision or accuracy, while measures based on the latter evaluate its bias. Our focus here is on measures of forecast accuracy.

Measures of accuracy commonly used to evaluate cross-sectional, subnational forecasts can be placed into one of two sets, those that are 'scale-dependent' and those that are not (Hyndman and Koehler 2006). Scale-dependent measures should be used with care when making accuracy comparisons across datasets so that different scales which affect the magnitude of these measures are not misinterpreted as differences in error. The most commonly used scale-dependent summary measures of forecast accuracy are based on the distributions of absolute errors (|E|) or squared errors (E^2), taken over the number of observations (n). These measures include:

$$\text{Mean Square Error (MSE)} = \left(\sum E^2\right)/n;$$

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\text{MSE}};$$

$$\text{Mean Absolute Error (MAE)} = \left(\sum |E|\right)/n; \quad \text{and}$$

$$\text{Median Absolute Error (MEDAE)} = \text{median}(|E|).$$

Both MSE and RMSE are integral components in statistical models (e.g. regression). As such, they are natural measures to use in many forecast error evaluations that use regression-based and statistical methods (Alho and Spencer 2005; Pflaumer 1988; Swanson 2008; Swanson and Beck 1994). There is no absolute criterion for a 'good' value of any of the scale-dependent measures. Moreover, as arithmetic means, the presence of outliers will influence MSE, RMSE, and MAE. Thus, they implicitly give greater weight to larger error values. One advantage that RMSE has over MSE is that its scale is the same as the forecast data. Instead of reporting in terms of the 'average' of squared errors, as is the case for MSE, errors reported by the RMSE are representative of the size of an 'average' error. MAE is also measured in the same units as the original data, and is usually similar in magnitude to, but slightly smaller than, the RMSE. MEDAE is not influenced by outliers, but this strength is also a weakness in that it does not maximize the use of available information on the errors, a trait it shares with many 'robust' measures.

Measures that are not scale-dependent adjust for the population size of the area using a percentage error given by $PE = (E/O) * 100$. Like the scale dependent

measures, a positive value of PE is derived by taking its absolute value (|PE|) or its square (PE^2). These measures include:

$$\text{Mean Square Percentage Error (MSPE)} = \left(\sum \text{PE}^2\right)/\text{n};$$

$$\text{Root Mean Square Percentage Error (RMSPE)} = \sqrt{(\text{MSPE})};$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \left(\sum |\text{PE}|\right)/\text{n}; \quad \text{and}$$

$$\text{MedianAbsolute Percentage Error (MEDAPE)} = \text{median}(|\text{PE}|).$$

Because percentage errors are not scale-independent, they are used to compare forecast performance across different datasets. The fact that they assume the existence of a meaningful zero is not a problem in demographic forecasting, as it would be if, for example, one were forecasting temperatures in the Fahrenheit or Celsius scales. However, they have a major disadvantage in that they are infinite or undefined if $O = 0$ for any observation. Moreover, because the underlying error distributions of these measures have only positive values and no upper bound, percentage errors are highly prone to right-skewed asymmetry (Smith and Sincich 1988). This means, for example, that the MAPE is often larger, sometimes substantially larger, than the MEDAPE. The MSPE and RMSPE provide the same properties as the MSE and RMSE, but are expressed as percentages.

The symmetrical MAPE (SMAPE) was designed to deal with some of the limitations of the MAPE (Makridakis 1993). Like MAPE, SMAPE is an average of the absolute percentage errors but these errors are computed using a denominator representing the average of the forecast and observed values. SMAPE has an upper limit of 200%, offers a well designed range to judge the level of accuracy and should be influenced less by extreme values. It also corrects for the computation asymmetry of the PE. For example, $F = 150$ and $O = 100$ yield a $PE = 50\%$, while $F = 100$ and $O = 150$ yield a $PE = 33\%$. The average of F and O in the denominator of the PE yields 40% in either situation. Despite these characteristics, the SMAPE is not a suitable alternative to the MAPE, and did not overcome its shortcomings (Tayman and Swanson 1999).

Other measures are based on relative errors (Armstrong and Collopy 1992; Hyndman and Koehler 2006; Swanson and Tayman 1995). These measures compare the accuracy from two forecasts, which can be based on different methods and assumptions. They can also compare a forecast from a naïve low-cost alternative to one based on a formal forecasting method. Measures based on relative errors are useful for judging the utility of a forecast, or its value in improving the quality of information upon which decisions are based. A summary of common summary error measures discussed below is presented along with their characteristics in Table 1.

Mean absolute percent error (MAPE)

Of the preceding measures, MAPE is most commonly used to evaluate cross-sectional, subnational forecasts (Ahlburg 1992, 1995; Campbell 2002; Hyndman

Table 1 Summary of common error measures and their characteristics

Measure	Name	Equation	Description
Absolute percent error	APE	|PE|	Absolute error for each observation
Mean absolute percent error	MAPE	$\sum$|PE|)/n	Mean of absolute errors
Absolute percent error transformed	APE-T	$(APE^{\lambda}) - \lambda/\lambda$ when $\lambda \neq 0$ or ln(APE) when $\lambda = 0$	Transformed absolute error for each observation
Mean absolute percent error transformed	MAPE-T	$\sum$|APE-T|)/n	Mean of transformed absolute errors
Mean absolute percent error re-expressed	MAPE-R	$[(\lambda)(\text{MAPE-T} + 1)]^{1/\lambda}$	MAPE-T Re-expressed into the original distribution scale

and Koehler 2006; Isserman 1977; Miller 2001; Murdock et al. 1984; Rayer 2007; Sink 1997; Smith 1987; Smith and Sincich 1990, 1992; Smith et al. 2001; Tayman et al. 1998; Wilson 2007). It is a signal of MAPE's ubiquity that it is often found in software packages such as Autobox, ezForecaster, Nostradamus, SAS, and SmartForecast. In addition, MAPE has valuable statistical properties in that it makes use of all observations and has the smallest variability from sample to sample (Levy and Lemeshow 1991). MAPE is also often useful for purposes of reporting, because it is expressed in generic percentage terms that will be understandable to a wide range of users.

MAPE is simple to calculate and easy to understand, which attest to its popularity, but does it meet the criteria for a good measure of error? According to the National Research Council (1980), any summary measure of error should meet five basic criteria: measurement validity, reliability, ease of interpretation, clarity of presentation, and support of statistical evaluation. MAPE meets most of these criteria, but its validity is questionable. As noted previously, the distribution of absolute percentage errors is often asymmetrical and right skewed. Thus the MAPE is neither a resistant nor a robust summary measure because a few outliers can dominate it and the MAPE will not be close in value for many distributions (Hoaglin et al. 1983, p. 28; Huber 1964; Tukey 1970). Therefore, the MAPE can understate forecast accuracy, sometimes dramatically. Consequently, it has tended to reinforce the perception of inaccurate forecasts.

Mean absolute percent error rescaled (MAPE-R)

That MAPE is subject to overstating error because of the presence of extreme outliers has long been known, and attempts to constrain the effect of outliers have taken several paths: (1) controlling variables like population size; (2) using a more resistant summary of the distribution like a median or M-estimators; or (3) trimming the tail of the distribution. However, as Swanson et al. (2000) argued, outliers do inform the improvement of population estimates and forecasts, which is the primary reason they introduced MAPE-Rescaled (MAPE-R). Eliminating outliers removes

information and MAPE-R was designed to preserve such information by 'normalization' rather than elimination. Two major advantages in using a transformed distribution are that all observations are kept in the analysis, and all measures of central tendency will be approximately the same if the transformed distribution is symmetrical. Among other things, in a symmetrical distribution the mean will be as robust and resistant as the median.

One might argue that having an upward bias in a summary measure of central tendency is desirable because large errors should be reflected, if present. We along with others (National Research Council 1980; Morrison 1971) take a different view. Measures of central tendency should not be the only criteria to evaluate the error in population projections. Large outlying errors can and should be examined separately from the central tendency of error. When data are symmetrically distributed the arithmetic mean provides the centre of gravity and the centre of probability, and characterizes the bulk of the distributions. The arithmetic mean provides only a centre of gravity when data are asymmetrical.

To change the shape of a distribution efficiently and objectively and to achieve parity for the observations, Swanson et al. (2000) use a standardized technique designed to generate a single, nonlinear function to change the shape of the APE distribution. This technique modifies the power transformation developed by Box and Cox (1964),[1] defined as:

$$y(\lambda) = \left(x^{\lambda} - \lambda\right)/\lambda \quad \text{when } \lambda \neq 0; \text{ or}$$
$$\mathrm{y}(\lambda) = \ln(\mathrm{x}) \quad \text{when } \lambda = 0,$$

where x is the absolute percentage error, y is the transformed observation, and λ is the power transformation constant. One determines Lambda (λ) by finding the λ value that maximizes the function:

$$ml(\lambda) = -(n/2) \times \ln\left[(1/n)\sum (y_i - \bar{y})^2\right] + (\lambda - 1) \times \sum \ln(x_i),$$

where n is the sample size; y is the transformed observation; $\bar{y}$is the mean of the transformed observations; x is the original observation.

According to Box and Cox (1964), ml(λ) at a local maximum provides the power transformation (λ) for x that optimizes the probability that the transformed distribution will be symmetrical. In other words, finding λ does not guarantee symmetry, but it represents the transformation power most likely to yield a symmetrical distribution. We can find the maximum value of ml(λ) by solving its function for different values of λ between the range of -2 and 2 and identifying the largest resulting Box-Cox value (Draper and Smith 1981, p. 225).

To address the effect of a skewed distribution on MAPE, Swanson et al. (2000) transformed the Absolute Percent Error (APE) distribution using a Box-Cox transformation and introduced MAPE-Transformed (MAPE-T) as a summary measure of accuracy for this transformed distribution. The transformed distribution considers the entire data series, but assigns a proportionate amount of influence to

[1] Swanson et al. (2000) used λ in the numerator. Box and Cox (1964) used 1.0 in their original development to assure continuity in λ when $\lambda = 0$. The difference is immaterial.

each case through normalization, thereby reducing the otherwise disproportionate effect of outliers on a summary measure of error.

In preliminary tests, Swanson et al. (2000) note that their modified Box-Cox transformation not only compressed very large values, but also increased values greater than one in skewed distributions where λ was relatively small (less than 0.4). This property illustrates why this transformation is more effective in achieving a symmetrical distribution than simpler, non-linear functions that only increase untransformed errors of less than one. Because many estimation errors are greater than one percent, the modified Box-Cox equation not only lowers extremely high values toward the body of the data, but also raises relatively low values. These characteristics minimize skewness and increase symmetry.[2]

The transformed APE distribution has a potential disadvantage: transformation may move the observations into a unit of measurement that is difficult to interpret (Emerson and Stoto 1983: 124). This is not a trivial issue. As mentioned earlier, the National Research Council (1980) states that an error measure must have clarity of presentation. It is easier to think of estimation error in terms of percentages than, for example, log-percentages or square root-percentages. Interpretation may be impeded if the modified Box-Cox transformation is used because it is even less intuitive than simpler transformations, such as the natural log and square root. In addition to reflecting a new unit of measurement, the average error of the transformed distribution may reflect a new scale that further complicates clear understanding and interpretation of error.

Tayman et al. (1999) suggest using one of two classes of nonlinear functions (quadratic and power) to re-express the scale of the transformed observations into the scale of the original observations. Using coefficients from regressions of the APEs on the APE-Ts, they solve for MAPE-R based on the value of MAPE-T. Initially, regression was considered an effective but cumbersome way to re-express MAPE-T into MAPE-R. Testing revealed a more serious problem. When λ approaches zero, regression results become inconsistent. With the closed form expression in mind (as well as the geometric mean), a simple procedure for re-expressing MAPE-T back into the original scale of MAPE was tested by Coleman and Swanson (2007).[3] This re-expression is found by taking the inverse of MAPE-T[4]:

[2] A potential shortcoming of the Box-Cox transformation is that it is not globally monotonic. Individual values may have differential influence on the function. Values near the mean of the transformed distribution have little effect, while extreme outliers may actually reduce the MAPE-T. Because the Box-Cox transformation has no associated influence function, it is difficult to determine if and when the Box-Cox will perform this way (Coleman and Swanson 2007).

[3] Coleman and Swanson (2007) find this closed form expression for MAPE-R to be a member of the family of power mean-based accuracy measures. This enables it to be placed in relation to other members of this family, which includes Harmonic Mean Absolute Percent Error (HMAPE), Geometric Mean Absolute Percent Error (GMAPE), and MAPE. Given that MAPE-R was designed to be robust in the face of outliers, it is not surprising to find that it is a valid estimator of the median of the distribution generating the absolute percentage errors. Simulation studies suggest that MAPE-R is a far more efficient estimator of this median than MEDAPE.

[4] If the optimal value of λ found by the Box-Cox procedure is small, between -0.4 and $+0.4$, the transformed APEs are sufficiently far from the original scale that re-expression is required (Tayman et al. 1999).

$$\text{MAPE-R} = [(\lambda)(\text{MAPE-T} + 1)]^{1/\lambda}.$$

Is MAPE-R needed?

Swanson et al. (2000) provide a set of guidelines for determining if MAPE-R is needed. The central issue is the symmetry in the distribution of APEs. If the distribution of APEs in a given forecast evaluation is symmetrical, then MAPE will appropriately reflect its centre of gravity. However, if it is right-skewed, with outliers in the upper tail, then the centre of gravity as measured by MAPE is vulnerable to being dominated by these outliers, which suggests that the APEs should be transformed into a more symmetrical distribution. In determining if a set of APEs should be so transformed, Emerson and Stoto (1983: 125) establish the following guideline: if the absolute ratio of the highest APE value to the smallest APE value exceeds 20, transformation may be useful; if the ratio is less than 2, then a transformation may not be useful; a ratio between 2 and 20 is indeterminate.

If the Emerson-Stoto guidelines find that a transformation is called for or if the question is indeterminate, Swanson et al. (2000) suggest using a statistical skewness test to make a final determination in regard to transformation of the APEs. We use the skewness test developed and tested by D'Agostino et al. (1990). The null hypothesis tested is that the skewness value = 0, using the 0.10 level of significance. We recommend this significance level rather than more stringent ones (e.g., 0.05 and 0.01) because there is a greater cost in terms of a downwardly biased measure of accuracy in not transforming a potentially skewed distribution.

When the guidelines indicate a potentially useful transformation of APEs to a symmetrical distribution, the transformation is assumed to be successful when the average of the new distribution does not overstate or understate the error level and uses all observations. In this situation, the observations receive nearly equal weights, closer to 1/n, while the resulting average remains intuitively interpretable and clear in its presentation.

Data

We conducted our analysis using a dataset covering all counties or county equivalents in the United States that did not experience significant boundary changes between 1900 and 2000 (Rayer 2008).[5] This dataset included 2,481 counties in 48 states, 79% of all counties. For each county, information was collected on population size in the launch year (the year of the most recent data used to make a forecast), growth rate over the base period (the 20 years immediately preceding the launch year), and forecast errors for 10- and 20-year horizons. The launch years included all decennial census years from 1920 to 1990. For this analysis, we selected a 2000 forecast derived from the 1970 to 1990 base period

[5] These data were kindly provided by Stefan Rayer, Bureau for Business and Economic Research, University of Florida.

(10-year horizon).[6] Forecast errors were calculated as the percentage difference between the population forecast in 2000 and the population counted in the 2000 decennial census. We refer to these differences as forecast errors, although they may have been caused partly by the errors in the census counts themselves.[7]

Forecasts were derived from five simple extrapolation techniques: linear, exponential, share of growth, shift share, and constant share (Rayer 2008). The forecasts analysed in this study were calculated as an average of the forecasts from these five techniques, after excluding the highest and lowest. Simple techniques such as these are frequently used for small-area forecasts and have been found to produce forecasts of total population that are at least as accurate as those produced using more complex or sophisticated techniques (Long 1995; Murdock et al. 1984; Smith and Sincich 1992; Smith et al. 2001). An important benefit of these techniques is that they rely on readily available data and can be applied easily to a very large data set. Given the similarity of errors for total population generally found for most forecasting techniques applied to the same geographic regions and time periods, we believe the results reported here are likely to be valid for other techniques and time periods as well.

Analysis

We begin by analysing the entire sample of 2,481 counties (see Table 2). The MAPE of the original APE distribution was 6.21%. The ratio of the highest to lowest APE (5,220) (Max/Min) indicates the need for transformation and the hypothesis of symmetry is rejected (*P*-value 0.000). Following Tayman et al. (1999), we use the ratios of the MAPE and the MAPE-R to MEDAPE, respectively, as indications of the bias of the average as a measure of accuracy.[8] For all counties, this ratio is 1.40 suggesting that MAPE understates the average forecast accuracy by 40%.

The Box-Cox transformation yields a λ value of 0.272 resulting in a MAPE-R of 4.42%, which is 29% less than the original MAPE.[9] The MAPE-R to MEDAPE

[6] This sample of 2,481 counties had average and median sizes in 1990 of 79,100 and 23,400 respectively. Between 1970 and 1990, they grew at average and median rates of 22 and 14.6%, respectively, with 29% showing population declines during this period. Rayer (2004) has shown that this restricted sample is representative of all U.S. counties.

[7] We did not adjust these results for changes in census coverage over time. Nationally, the net census undercount has declined since 1950, except for a small increase between 1980 and 1990. In 2000, both demographic analysis and postenumeration surveys showed a slight overcount at the national level (Robinson et al. 2002; U.S. Census Bureau 2003). To our knowledge, estimates of census coverage errors for counties do not exist. Although changes in census coverage undoubtedly had some effect on the result reported here, we believe these effects are relatively small and do not affect our findings.

[8] This ratio is used as a descriptive tool to help judge the influence of outliers on the MAPE. We relate two measures of accuracy; one, MAPE, is affected by outliers and the other, MEDAPE, is not. In this application, MEDAPE is a convenient reference point that provides an error measure free of the influence of extreme values.

[9] The transformation of very small original APEs (<0.20%) resulted in slightly negative values in four counties (0.2%). Negative values do not substantially affect the resulting MAPE-R and should be set to zero in practice.

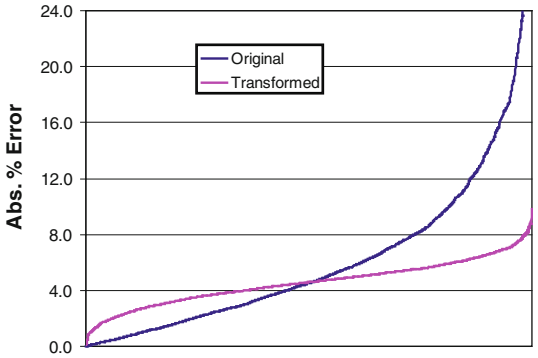

Fig. 1 Absolute percent error, U.S. Counties

Table 2 APE distribution statistics, all counties

	Untransformed	Transformed
Sample size	2,481	2,481
Lambda	n/a	0.272
Skewness	2.099	−0.017
P-value[a]	0.000	0.677
Max/Min[b]	5,220	196
MAPE	6.21	n/a
MAPE-R	n/a	4.42

[a] Ho: Skew = 0

[b] Maximum APE/Minimum APE

ratio drops to 0.98, indicating that the transformed APE distribution is much less influenced by outlying errors than the original APE distribution. The Box-Cox normalization is successful in that the transformed APE distribution has a skewness coefficient close to zero and the null hypothesis of symmetry is accepted. The Max/Min ratio has decreased substantially to 196; however, if the 10 values of less than 0.5 are excluded that ratio drops to 19.6.

The effect of the transformation is seen in Fig. 1, which compares the original APE and transformed APE-T distributions for all counties. The transformation modestly increases the APEs up to the median of the distribution where APE and APR-T are roughly equal. Values above the median are adjusted downward at an increasing rate, which applies the largest adjustments to the most extreme values. As a result, the transformed distribution is no longer influenced by outlying errors and the resulting MAPE-R is smaller than the MAPE.[10]

The usefulness of the Box-Cox normalization and resulting MAPE-R as an average measure of forecast accuracy not influenced by outliers is evident when looking at all counties. These results are in line with results reported previously. We

[10] Coleman (personal communication 2010) suspects the inverse correlation between lambda and the percentage reduction in average error may be perfect and nonlinear. He points out that when lambda = 1, one is using the data 'as is,' while when lambda = 0, one is 'squeezing' the data by taking logarithms. Consequently, higher values are reduced more than lesser ones, while intermediate values of lambda produce intermediate amounts of 'squeezing' which increases as lambda goes to 0.

Table 3 APE transformation decision, states

	No.	Percent
Insufficient Sample Size	6	12
No Transformation Suggested	7	15
Transformation Suggested	35	73
	48	100

now take the analysis a step further by examining the county errors separately for the 48 contiguous states in our sample. We first assess the original APE distribution for the counties in each state to determine whether or not a transformation is warranted. Table 3 provides a summary of the transformation decision and detailed results are found in Table 4.

Eight observations are required to run the skewness test (D'Agostino et al. 1990), which eliminates six states (12%) from further analysis. In a skewed distribution with very small samples, the median is a robust measure of central tendency that would not discard much information.

A transformation was not suggested in seven states (15%). The hypothesis of symmetry was accepted with *P*-values greater than 0.10 in these states, and in five out of seven the Max/Min ratio was less than 20. The Max/Min ratio was greater than 20 in New Jersey and South Carolina. In New Jersey, excluding the one value less than 1.5% reduces the ratio to 4.0, while for South Carolina; excluding the 3 values less than 0.3 reduces it to 19.9.

We would expect the MAPE and MAPE-R to be similar in these seven states. This occurs for the counties in Connecticut, Maine, New Hampshire, and South Carolina. MAPE-R is within 10% of the MAPE for each state as shown in the last column of Table 4. Counties in Arizona, New Jersey, and Vermont show a greater divergence between the MAPE and MAPE-R, suggesting a transformation. An ascending sort of the APEs in these states shows a nonlinear pattern, causing an upward bias to their MAPEs (data not shown). The MAPE to MEDAPE ratio in these states ranged from 1.17 to 1.28. The ratio in the four states where the MAPE and MAPE-R were similar ranged from 0.96 to 0.99. The sample sizes in Arizona (11), New Jersey (21), and Vermont (14) indicate that the skewness test may not pick up modest departures from symmetry in small samples.

Transformations were suggested for 35 of the 48 states (73%) by both the Max/Min ratio and skewness test. In these states, the Max/Min ratio was greater than 20 and ranges from 39 to over 11,000. The hypothesis of no skewness was rejected in each state at the 0.10 level. Across these states the average of the MAPE/MEDAPE ratios is 1.32, suggesting on average a 32% understatement of forecast accuracy in the original APE distributions. The Box-Cox normalization appears to work well in states whose counties show a skewed APE distribution. The skewness values of the transformed APEs are close to zero (ranging from −0.196 to 0.030) and the hypothesis of no skewness is accepted in each state. The MAPE-R is less than the MAPE for every state, reducing the average error (i.e., increasing the degree of accuracy) by an average of 26%, with a range of 9.2% to 60.6% (last column in

Table 4 APE distribution statistics, 48 contiguous states

State	Sample Size	Untransformed				Transformed				% Reduction[c] in average error
		MAPE	Skew	*P*-value[a]	Max/Min[b]	Lambda	MAPE-R	Skew	*P*-value[a]	
Insufficient sample size										
Delaware	3	8.09	1.150	n/a	17	n/a	n/a	n/a	n/a	n/a
Idaho	4	8.05	1.302	n/a	6	n/a	n/a	n/a	n/a	n/a
Montana	5	11.01	1.453	n/a	7	n/a	n/a	n/a	n/a	n/a
New Mexico	4	7.74	0.894	n/a	4	n/a	n/a	n/a	n/a	n/a
Rhode Island	3	1.99	2.812	n/a	5	n/a	n/a	n/a	n/a	n/a
Wyoming	4	6.77	0.802	n/a	265	n/a	n/a	n/a	n/a	n/a
Transformation not suggested										
Arizona	11	9.96	0.589	0.358	9	0.090	7.700	−0.043	0.946	22.7%
Connecticut	8	2.80	−0.108	0.884	3	0.644	2.700	−0.219	0.768	3.7%
Maine	12	8.64	0.051	0.934	8	0.826	8.470	−0.149	0.805	2.0%
New Hampshire	10	4.30	−0.065	0.922	9	0.656	4.050	−0.265	0.689	5.8%
New Jersey	21	5.48	0.499	0.300	50	0.458	4.740	−0.031	0.778	13.6%
South Carolina	18	4.86	−0.067	0.849	60	0.671	4.400	−0.457	0.372	9.4%
Vermont	14	4.06	0.904	0.125	17	0.227	3.280	−0.069	0.901	19.2%
Transformation suggested										
Alabama	59	4.69	1.228	0.001	157	0.338	3.613	−0.083	0.781	22.9%
Arkansas	62	5.35	1.322	0.000	158	0.314	3.900	−0.093	0.742	27.1%
California	54	6.07	0.836	0.014	255	0.340	4.784	−0.094	0.757	21.2%
Colorado	48	18.89	0.735	0.036	562	0.538	17.150	−0.053	0.872	9.2%
Florida	26	6.92	1.832	0.001	39	0.147	5.104	−0.021	0.956	26.2%
Georgia	101	9.99	1.423	0.000	3,738	0.320	7.244	−0.068	0.760	27.5%
Illinois	102	4.36	1.753	0.000	156	0.194	3.086	−0.029	0.893	29.2%

Table 4 continued

State	Sample Size	Untransformed				Transformed				% Reduction[c] in average error
		MAPE	Skew	*P*-value[a]	Max/Min[b]	Lambda	MAPE-R	Skew	*P*-value[a]	
Indiana	92	4.68	2.148	0.000	11,076	0.407	3.632	−0.083	0.720	22.4%
Iowa	99	4.83	1.625	0.000	77	0.376	4.097	−0.028	0.897	15.1%
Kansas	105	4.66	1.127	0.000	161	0.372	3.783	−0.077	0.729	18.9%
Kentucky	116	6.32	1.887	0.000	3,418	0.335	4.653	−0.052	0.800	26.4%
Louisiana	56	4.98	2.430	0.000	665	0.384	3.939	0.030	0.932	20.9%
Maryland	19	4.01	1.457	0.009	64	0.202	2.785	−0.060	0.899	30.6%
Massachusetts	10	4.78	2.761	0.000	215	0.052	1.884	−0.028	0.962	60.6%
Michigan	83	5.15	2.944	0.000	116	0.090	3.445	−0.009	0.965	33.1%
Minnesota	78	5.78	1.530	0.000	170	0.240	4.364	−0.006	0.974	24.5%
Mississippi	60	5.85	2.025	0.000	5,816	0.303	3.800	−0.081	0.775	35.0%
Missouri	115	5.69	1.453	0.000	92	0.272	4.396	−0.049	0.814	22.7%
Nebraska	87	5.74	1.221	0.000	177	0.405	4.737	−0.054	0.820	17.4%
Nevada	9	11.52	1.418	0.050	39	0.096	6.631	−0.050	0.941	42.5%
New York	56	3.47	1.616	0.000	474	0.304	2.423	−0.096	0.745	30.1%
North Carolina	86	7.42	0.965	0.001	3,702	0.465	6.331	−0.084	0.729	14.7%
North Dakota	31	4.99	0.966	0.026	2,733	0.309	3.162	−0.196	0.619	36.6%
Ohio	88	3.39	3.537	0.000	265	0.133	2.070	−0.009	0.959	38.9%
Oklahoma	9	4.23	2.245	0.003	105	0.206	2.722	−0.018	0.975	35.7%
Oregon	29	6.46	1.414	0.003	58	0.186	4.481	−0.052	0.895	30.6%
Pennsylvania	67	3.36	2.812	0.000	311	0.110	2.039	−0.015	0.951	39.3%
South Dakota	49	7.50	1.121	0.003	94	0.443	6.168	−0.143	0.654	17.8%
Tennessee	94	7.15	0.582	0.022	277	0.500	6.178	−0.145	0.544	13.6%
Texas	232	7.94	1.133	0.000	258	0.279	5.733	−0.074	0.628	27.8%

Table 4 continued

State	Sample Size	Untransformed				Transformed				% Reduction[c] in average error
		MAPE	Skew	*P*-value[a]	Max/Min[b]	Lambda	MAPE-R	Skew	*P*-value[a]	
Utah	25	12.81	0.804	0.082	224	0.469	11.000	−0.066	0.875	14.1%
Virginia	66	7.79	1.943	0.000	48	0.174	6.044	−0.015	0.952	22.4%
Washington	32	7.38	0.789	0.059	42	0.361	6.318	−0.031	0.932	14.5%
West Virginia	55	4.31	1.772	0.000	214	0.226	2.991	−0.048	0.869	30.5%
Wisconsin	64	4.45	0.713	0.021	95	0.519	3.947	−0.107	0.705	11.3%

[a] Ho: Skew = 0

[b] Maximum APE/Minimum APE

[c] (1 − (MAPE/MAPE-R)) * 100

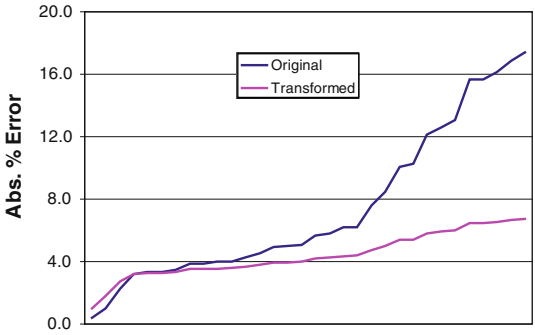

Fig. 2 Absolute percent error, Washington Counties

Table 4). Moreover, the average of the MAPE-R/MEDAPE ratios is now 0.98, indicating the success of the transformation of achieving measures of average error not influenced by outliers.

To measure the impact of the skewness on the upward bias of the MAPE, we regressed the percentage reduction in error against the skewness in the original APE distribution for the 35 states where a transformation was suggested. A power function using the natural log of both variables best describes this relationship with an adjusted R^2 of 0.508. The elasticity coefficient (0.670) indicates that for every 1% increase in skewness the upward bias of the MAPE increases by approximately 0.7%.[11] Before this research, the sample sizes were insufficient to estimate this effect.

A possible concern with this transformation is that it could make the data look 'better' than they really are. However, the re-expression of the original APE can either increase or decrease the MAPE-T relative to the MAPE. The latter will usually occur with more frequency. In the 35 states, the MAPE-T is smaller than the MAPE in 26 and larger in only nine. To illustrate the effect of the transformation for these conditions we examined the error by county in Washington (MAPE = 7.38% and MAPE-T = 4.32) and Pennsylvania (MAPE% = 3.36 and MAPE-T = 8.83).

Figure 2 shows the more typical case illustrated by results for the State of Washington, where a λ of 0.361 results in an upward adjustment of a relatively few small original APEs, modest downward adjustments to errors close to the main body of the data, and an increasing downward adjustment as the APE moves away from the bulk of the observations. In this case, the MAPE-T is portraying an average accuracy that is probably biased downward; therefore the adjustment to MAPE-R increases the average to 6.32%. In Pennsylvania, a different pattern of adjustment emerges (see Fig. 3). The smaller λ of 0.110 causes the bulk of the observations to adjust upwards. This upward adjustment lessens toward the upper end of the distribution and for the relatively few highest values a substantial downward adjustment occurs. The average of the transformed APE (MAPE-T) is substantially

[11] A regression model including Arizona, New Jersey, and Vermont had an adjusted R^2 of 0.494 and an elasticity coefficient of 0.508, showing a slightly smaller impact of skewness on upward bias.

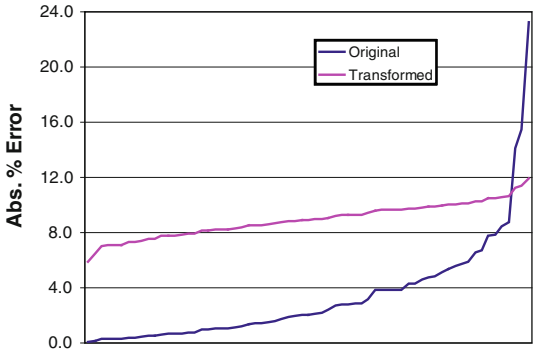

Fig. 3 Absolute percent error, Pennsylvania Counties

larger than the MAPE and is probably biased upward; therefore the adjustment to MAPE-R decreases that average considerably to 2.04%.

Conclusions and suggestions for future research

MAPE is a suitable measure in many instances (Campbell 2002; Rayer 2007). However, it is often based on a right skewed distribution of APEs, which pulls the average error upward and understates the forecast accuracy of the bulk of the observations. Releasing evaluations that understate accuracy only serves to perpetuate the perception that demographic forecasts are inaccurate.

To determine if average error is overstated, we offered a two-step process for evaluating the shape of the original APE distribution. The first step is the Emerson-Stoto test (Emerson and Stoto 1983: 125), which is based on the ratio of the maximum to minimum APE. If the first step suggests that the distribution of APEs does require transformation, use a formal hypothesis test of skewness (e.g., D'Agostino et al. 1990) to make the final decision. We also showed how these criteria can be used to judge the symmetry of the APE distribution, instead of simply accepting MAPE (or MEDAPE) in a given situation. If these tests indicate that MAPE is not suitable, we advise using MAPE-R. When a transformation is indicated by the two-step process, we believe MAPE-R represents an improvement over MAPE in evaluating the accuracy of cross-sectional, subnational population forecast as well as cross-sectional, subnational population estimates. Moreover unlike MEDAPE, MAPE-R preserves information about the structure of error in the presence of the outliers that affect MAPE.

We also demonstrated that the Box-Cox transformation can normalize a skewed distribution of APEs and, further, that a simple procedure to re-express the distribution average (MAPE-T) provided a summary measure of error (MAPE-R) that is robust, resistant, and compliant with the standards set by the National Research Council (1980). This procedure worked over a wide range of APE distributions and accommodated situations where original APEs were predominately adjusted either up or down and where the skewness ranged from moderate to extreme. In sum, using MAPE-T retains all of the information in the forecast error

distribution and reduces the effect of outliers on the summary measure. Using the technique described by Coleman and Swanson (2007), MAPE-R is easy to calculate from MAPE-T, is more consistent in terms of monotonicity, and is readily understandable. Most important MAPE-R is an average measure of forecast accuracy that is not influenced by the relatively few large errors that tend to characterize the distribution of APEs.

This empirically-based examination of MAPE-R suggests four areas of future research:

1. Improving the availability of MAPE-R by developing software that can be easily accessed in a user-friendly computing environment.
2. Exploring the effect on MAPE-R using an omnibus test of normality that includes both skewness and kurtosis (e.g., Jarque and Bera 1987) instead of a test that just considers skewness.
3. Investigating normality functions with influence curves to address the problems of non-global monotonicity and known instability of the Box-Cox transformation. One possibility is the geometric average (GMAPE), which like MAPE-R is not subject to the shortcomings of the MAPE or to the instability of the Box-Cox transformation.
4. Examining the sensitivity of the λ calibration and the structure of the transformations. For example, would the results by state differ if based on a single calibration of all counties rather than a state by state calibration as shown in this paper?

In this paper, we have provided a large scale empirical test of MAPE-R and a set of refined guidelines for its use. Evidence in the test and elsewhere suggests that cross-sectional, subnational forecasts are subject to errors that often include substantial outliers. The results suggest that MAPE-R should be used instead of MAPE in evaluating these forecasts if substantial outliers are indicated through the guidelines we offer. Moreover, we argue that unlike MEDAPE, MAPE-R preserves useful information about the structure of error in the presence of the substantial outliers. Thus the MAPE-R offers a more meaningful representation of average error than either MAPE or MEDAPE when evaluations of cross-sectional, subnational population forecasts indicate that substantial outliers are present.

References

Ahlburg, D. (1992). A commentary on error measures: Error measures and choice of a forecast method. *International Journal of Forecasting, 8*, 99–111.

Ahlburg, D. (1995). Simple versus complex models: Evaluation, accuracy, and combining. *Mathematical Population Studies, 5*, 281–290.

Alho, J., & Spencer, B. (2005). Statistical demography and forecasting. In W. Alonso & P. Starr (Eds.), *The politics of numbers*. New York: Russell Sage.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*, 69–80.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B, 26*, 211–252.

Campbell, P. (2002). *Evaluating forecast error in state population projections using Census 2000 counts.* Population Division Working Paper Series No. 57. Washington, D.C.: U. S. Census Bureau.

Coleman, C., & Swanson, D. (2007). On MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement, 32*(4), 219–233.

D'Agostino, R., Belanger, A., & D'Agostino, R. Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(3), 316–321.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: John Wiley.

Emerson, J., & Stoto, M. (1983). Transforming data. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 97–128). New York: Wiley.

Hoaglin, D., Mosteller, F., & Tukey, J. (1983). Introduction to more refined estimators. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 283–296). New York: Wiley.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73–101.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679–688.

Isserman, A. (1977). The accuracy of population projections for subcounty areas. *Journal of the American Institute for Planners, 43*, 247–259.

Jarque, C., & Bera, A. (1987). A test for normality of observations and regression residuals. *International Statistical Review, 55*(2), 163–172.

Levy, P., & Lemeshow, S. (1991). *Sampling of populations: Methods and applications*. New York: Wiley.

Long, J. (1995). Complexity, accuracy, and utility of official population projections. *Mathematical Population Studies, 5*, 203–216.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting, 9*, 527–529.

Miller, E. R. (2001). *Evaluation of the 1990 school district level population estimates based on the synthetic ratio approach*. Population Division Working Paper No. 54. Washington D.C.: US Census Bureau.

Morrison, P. (1971). *Demographic information for cities: A manual for estimating and projecting local population characteristics*. Santa Monica, CA: Rand.

Mulder, T. (2002). *Accuracy of the U.S. Census Bureau national population projections and their respective components of change*. Population Division Working Paper No. 50. Washington, D.C.: U.S. Census Bureau. (http://www.census.gov/population/www/documentation/twps0050/twps0050.html, last accessed January 2010).

Murdock, S., Leistritz, L., Hamm, R., Hwang, S., & Parpia, B. (1984). An assessment of the accuracy of a regional economic-demographic projection model. *Demography, 21*, 383–404.

National Research Council. (1980). *Estimating population and income for small places*. Washington, D.C.: National Academy Press.

Pflaumer, P. (1988). Confidence intervals for population projections based on Monte Carlo methods. *International Journal of Forecasting, 4*, 135–142.

Rayer, S. (2004). *Assessing the accuracy of trend extrapolation methods for population projections: The long view*. Paper presented at the annual meeting of the Southern Demographic Association. Hilton Head Island, South Carolina.

Rayer, S. (2007). Population forecast accuracy: Does the choice of summary measure of error matter? *Population Research and Policy Review, 26*, 163–184.

Rayer, S. (2008). Population forecast errors: A primer for planners. *Journal of Planning Education and Research, 27*, 417–430.

Robinson, J., West, K., & Adlakha, A. (2002). Coverage of the population in Census 2000: Results from demographic analysis. *Population Research and Policy Review, 21*, 19–38.

Sink, L. (1997). *Race and ethnicity classification consistency between the Census Bureau and the National Center for Health Statistics*. Population Division Working Paper No. 17. Washington, D.C.: U.S. Census Bureau.

Smith, S. (1987). Tests of accuracy and bias for county population projections. *Journal of the American Statistical Association, 82*, 991–1003.
Smith, S., & Sincich, T. (1988). Stability over time in the distribution of population forecast errors. *Demography, 25*, 461–474.
Smith, S., & Sincich, T. (1990). On the relationship between length of base period and population forecast errors. *Journal of the American Statistical Association, 85*, 367–375.
Smith, S., & Sincich, T. (1992). Evaluating the forecast accuracy and bias of alternative projections for states. *International Journal of Forecasting, 8*, 495–508.
Smith, S., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.
Swanson, D. A. (2008). Measuring uncertainty in population data generated by the cohort-component method: A report on research in progress. In S. Murdock & D. A. Swanson (Eds.), *Applied demography in the 21st Century* (pp. 165–189). Dordrecht: Springer.
Swanson, D. A., & Beck, D. (1994). A new short-term county population projection method. *Journal of Economic and Social Measurement, 20*, 1–26.
Swanson, D. A., & Stephan, G. E. (2004). Glossary and demography timeline. In J. Siegel & D. A. Swanson (Eds.), *The methods and materials of demography* (2nd ed., pp. 751–786). San Diego, CA: Elsevier Academic Press.
Swanson, D. A., & Tayman, J. (1995). Between a rock and a hard place: The evaluation of demographic forecasts. *Population Research and Policy Review, 14*, 233–249.
Swanson, D. A., Tayman, J., & Barr, C. F. (2000). A note on the measurement of accuracy for subnational demographic estimates. *Demography, 37*, 193–201.
Tayman, J., Schafer, E., & Carter, L. (1998). The role of population size in the determination of population forecast errors: An evaluation using confidence intervals for subcounty areas. *Population Research and Policy Review, 17*, 1–20.
Tayman, J., & Swanson, D. A. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review, 18*, 299–322.
Tayman, J., Swanson, D. A., & Barr, C. F. (1999). In search of the ideal measure of accuracy for subnational demographic forecasts. *Population Research and Policy Review, 18*, 387–409.
Tukey, J. (1970). *Exploratory data analysis* (Vol. I). Reading, MA: Addison-Wesley.
U.S. Census Bureau (2003). Technical assessment of the A.C.E. revision II. Available on-line at http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf.
Wilson, T. (2007). The forecast accuracy of Australian Bureau of Statistics national population projections. *Journal of Population Research, 24*(1), 91–117.
Yokum, J., & Armstrong, J. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting, 11*, 591–597.