Small-Area Population Forecasting Using a Spatial Regression Approach

Guangqing Chi and Paul R. Voss
Applied Population Laboratory
Department of Rural Sociology
University of Wisconsin-Madison

Extended abstract

The use of familiar multiple regression models in the production of population forecasts for subcounty geographic entities is an empirically tested idea that has been around for more than 50 years (Schmitt, 1953[1]; 1954[2]). Stanbery (1952)[3] did not mention regression-based forecasts in this early “guide book” for those responsible for population forecasting for small areas and communities. But, nearly a quarter century later, Pittinger (1976)[4], in his comprehensive review of population projection models devoted considerable attention to the matter (1976:68-77). In their recent and well-received overview on the topic, Smith, Tayman and Swanson (2000)[5] devote two chapters to structural modeling (of particular relevance to us in this paper is their Chapter 9 dealing with economic-demographic structural models), yet it is our view that most applied demographers making population forecasts for small areas (including the second author on this paper) have largely ignored regression forecasting approaches.

In this paper, we explore some of the likely reasons for declining interest in regression forecast models – chiefly among these the belief that they simply do not perform very well. We examine some of the alternatives for small area forecasting (such as temporal extrapolation models – which also do not have a distinguished track record). And, finally, we introduce, and formally test, a revised regression specification that brings into the regression forecasting

[1] Schmitt, R.C. 1953. "A New Method of Forecasting City Population." *Journal of the American Institute of Planners* 19(1):40-42.
[2] Schmitt, R.C. 1954. "A Method of Projecting the Population of Census Tracts." *Journal of the American Institute of Planners* 20(2):102.
[3] Stanbery, V.B. 1952. *Better Population Forecasting For Areas and Communities: A Guide Book for Those Who Make or Use Population Projections*. Washington D. C.: Superintendent of Documents, U.S. Government Printing Office.
[4] Pittenger, D.B. 1976. *Projecting State and Local Populations*. Cambridge, MA: Ballinger Publishing Company.
[5] Smith, S.K., J. Tayman, and D.A. Swanson. 2000. *State and Local Population Projections: Methodology and Analysis*. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic/Plenum Publishers.

approach explicit “neighborhood” influences through spatial regression (spatial econometric) techniques.

To begin, we define what we mean by a regression forecasting model. Using matrix notation, the standard multiple regression model is expressed as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad [1]$$

where: $\boldsymbol{y}$, the dependent variable, is a (n x 1) vector of realizations of a random variable,

$\boldsymbol{X}$ is a (n x k) matrix of fixed observations on independent variables,

$\boldsymbol{\beta}$ is a (k x 1) vector of parameters, and

$\boldsymbol{\varepsilon}$ is a (n x 1) vector of error terms.

A multiple regression forecasting model proceeds in two steps. In step one, it takes some function f($\boldsymbol{y}$) of population change, and establishes a relationship between f($\boldsymbol{y}$) and the variables in the design matrix $\boldsymbol{X}$, which are carefully chosen covariates of f($\boldsymbol{y}$), such that unbiased and efficient estimates of the vector of parameters can be achieved using the least squares estimator:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}\ \boldsymbol{X}\right)^{-1}\boldsymbol{X}\ \boldsymbol{y} \qquad [2]$$

By “carefully chosen” we mean that the independent variables are not highly intercorrelated, that they each are approximately linearly related to the dependent variable and that all variables are reasonably bell-shaped. The model must meet several other rather strict assumptions in order that the least squares estimates are unbiased and efficient. These are clearly spelled out in the standard econometrics literature (see, for example, Fox, 1997[6]; Draper and Smith, 1998[7]; Greene, 2000[8]). Our independent variables are fixed observations at time t, or observations over the interval (t-10, t), and the dependent variable is an observation for the period (t, t+10). In other words, our regression model establishes a set of relationships between

[6] Fox, J. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: SAGE Publications.

[7] Draper, N.R.and H. Smith. 1998. *Applied Regression Analysis*. New York, NY: John Wiley &Sons, Inc.

[8] Greene, W.H. 2000. *Econometric Analysis*. Upper Saddle River, NJ: Prentice-Hall, Inc.

growth over a decade (the dependent variable) with the growth over the preceding decade and several initial conditions at time t.

In step two, we use the relationships established in step 1, and expressed in the estimates of the β vector, to update the independent variables for the decade (t, t+10), and other initial conditions at time t+10 to forecast population change over the decade (t+10, t+20). The critical assumption is that the relationships between the independent variables and dependent variable, established in the base period, remain relatively constant over time and can thus be used to forecast change in a decade ten years beyond the base period where the relationships (the estimated β vector) were established.

It is important to point out that we deliberately are excluding from our specification of regression forecasting models those methods based on trend modeling (linear or quadratic regressions fit on a historical time series) as well as those methods based on adaptive smoothing, Box-Jenkins ARIMA modeling and the many related time-series approaches (see, for example, Thomopoulos, 1980[9]; Box and Jenkins, 1976[10]; McCleary and Hay, 1980[11]). These approaches have been proposed and effectively used for population forecasting (Alho and Spencer, 1997[12]; Pflaumer, 1992[13]; Saboia, 1974[14]), although the lack of systematic regularity in population time-series generally yields little beyond a trend forecast (and confidence intervals of dubious practical value).

Also excluded for further consideration in this paper are the post-censal population estimation models (e.g., the ratio-correlation method) that rely on contemporaneous systematic indicators for the estimate. The literature covering this regression approach to post-censal estimation is large (see, for example, Feeney, Hibbs, and Gillaspy, 1995[15]).

[9] Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

[10] Box, G.E.P.and G.M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

[11] McCleary, R.and R.A. Hay. 1980. *Applied Time Series Anslysis for the Social Sciences*. Landon, England: Sage Publications.

[12] Alho, J.and B. Spencer. 1997. "The Practical Specification of the Expected Error of Population Forecasts." *Journal of Official Statistics* 13:203-225.

[13] Pflaumer, P. 1992. "Forecasting U.S. Population Totals with the Box-Jenkins Approach." *International Journal of Forecasting* 8:329-338.

[14] Saboia, J. 1974. "Modeling and Forecasting Populations by Time Series: The Swedish Case." *Demography* 11:483-492.

[15] Feeney, D., J. Hibbs, and R.T. Gillaspy. 1995. "Ratio-Correlation Method." in *Basic Methods for Preparing Small-Area Population Estimates*, edited by N.W. Rivers, W.J. Serow, A.S. Lee, H.F. Goldsmith, and P.R. Voss. Madison, WI: University of Wisconsin-Madison/Extension.

On a more positive note, we draw attention to the improvements we make in this analysis to regression forecasting. We begin with a standard regression approach to modeling population change and then modify the regression specification to include explicit spatial spillover effects using the tools of spatial econometrics (Anselin, 1988[16]). We argue that nearly all forecasting models for small areas, even though widely disparate in their individual methodologies, share a single common shortcoming. They treat each unit of geography (e.g., a census tract, a minor civil division, a small city) as an independent, stand alone entity rather than as an entity surrounded by other geographic areas with which they interact (e.g., through commuting patterns, shopping patterns, etc.). We further argue that techniques now exist, although they have not found their way into population forecasting models to include "neighborhood effects" in regression models such that the forecast for entity A explicitly recognizes the forecast (and possibly other covariates) for neighboring entities B, C, D, etc.

In addition, we take regression model forecasting in new directions by including in our design matrix, ***X***, a number of nontraditional variables. A major shortcoming of existing regression forecasts is that they generally ignore non-demographic factors that are part of the contextual region for which the population forecasts are made. Population growth or decline has causes and consequences closely tied to levels of economic development as well as the nature of the surrounding natural environmental. Yet these latter factors are typically brushed aside when demographic change modeled. These factors can and should be incorporated into multivariate regression models for population forecasting.

In this study, the relevant data will be brought together using GIS tools. A spatial regression model will be applied to estimate the covariates of all related variables and population change, and the covariates will then be used for the population forecast. The central research question is whether this approach is effective for small-area population forecast when compared to standard methodologies now in practice.

The context for the study and the data come by examining population change at the Minor Civil Division (MCD) level in Wisconsin since 1970. A spatial lag model will be applied in this analysis for its strengths in considering neighbor effects, population spillovers, etc. For each MCD, in the first stage of the forecasting process, the population growth rate for 1980-1990

[16] Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

(the dependent variable) is regressed on several variables covering the 1970-1980 period and stock measures from the 1980 Census. Ultimately this model is expanded to include neighborhood characteristics in 1980, and neighborhood growth rates for 1970-1980. The characteristics include demographic, environmental, and socio-economic characteristics. The regression coefficients ($\boldsymbol{\beta}$'s) and a spatial regression parameter (λ) are estimated from this model, and these are used, in the second stage, to forecast the population growth rate in 1990-2000. The population forecasts for 2000, thus derived, are compared to the actual population in 2000 to calculate the Mean Algebraic Percent Error (MALPE) and the Mean Absolute Percent Error (MAPE). Similar error measures will also be calculated for the 2000 population projections earlier made by demographers in the Wisconsin state demographic agency, the Demographic Services Center.