Bode, Eckhardt

Working Paper

Annual educational attainment estimates for US counties 1990 - 2005

Kiel Working Paper, No. 1665

Provided in Cooperation with:
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Bode, Eckhardt (2010) : Annual educational attainment estimates for US counties 1990 - 2005, Kiel Working Paper, No. 1665, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/43122

Kiel
Working Papers

Kiel Institute
for the World Economy

Annual Educational Attainment Estimates for US Counties 1990 – 2005

by Eckhardt Bode

No. 1665 | November 2010

Web: www.ifw-kiel.de

Kiel Working Paper No. 1665 | November 2010

Annual Educational Attainment Estimates for US Counties 1990 -- 2005

Eckhardt Bode

Abstract:
This paper estimates annual data on educational attainment for 3,076 mainland U.S. counties 1991 -- 2005. Being estimated without resorting to ancillary information, this data is suited particular well for panel regression analyses. Several plausibility checks indicate that the data is fairly reliable and yields plausible parameter estimates in a panel regression.

Keywords: Educational attainment, U.S. counties, Panel regression

JEL classification: C33, C61, R12

Eckhardt Bode
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone: +49 (431) 8814-462
E-mail: eckhardt.bode@ifw-kiel.de

Annual Educational Attainment Estimates for U.S. Counties 1990 – 2005

Eckhardt Bode

November 24, 2010

Abstract

This paper estimates annual data on educational attainment for 3,076 mainland U.S. counties 1991 – 2005. Being estimated without resorting to ancillary information, this data is suited particular well for panel regression analyses. Several plausibility checks indicate that the data is fairly reliable and yields plausible parameter estimates in a panel regression.

Keywords: Educational attainment, U.S. counties, Panel regression
JEL: C33, C61, R12

1 Introduction

Advances in panel data regression techniques and the increasing availability of space-time panel data have facilitated controlling for unobserved, time-invariant factors in empirical studies of spatial phenomena. This helps reduce the biases of estimators that may arise from pure cross-section regressions. For a variety of economic indicators, annual data is, however, available only for more recent years. For earlier years, when panel data regression techniques had not been available or had not been employed frequently, this data is available only for selected years. To effectively use panel data regression techniques, filling in these gaps in data availability by estimating or interpolating the missing data is helpful.

The present paper estimates data on educational attainment of residents aged 25 or more in 3,076 mainland U.S. counties during the period 1991 – 2005. Population is divided into three exhaustive and mutually exclusive groups: (i) persons holding a bachelor degree or higher, (ii) persons holding a high-school diploma or higher but no bachelor degree, and (iii) persons with less than a high-school diploma. This data is, at the county level, available only from the decennial censuses, i.e., for every tenth year (e.g., 1990, 2000). In its American Community Survey, the United States Census Bureau (USCB) has recently started publishing own annual estimates of educational attainment for selected

counties from 2001 onwards.[1] Even though the number of counties for which the USCB published annual estimates has increased considerably over time (21 in 2001, 792 in 2009), this database is still far from being comprehensive. By estimating—and making publicly available—this data for the years 1991 to 1999 and 2001 to 2005, this study facilitates a significant expansion of the time dimension available for county-level panel regression analyses of economic phenomena related to the educational composition or human-capital intensity of the regional populations. To maximize the scope of economic analyses for which this data can be used, we explicitly refrain from using ancillary information in our estimations of this data. We estimate educational attainment only from the available data on educational attainment and the corresponding population totals. We use no further information on the compositions of the county populations. We even do not use geographical information such as distances or spatial weights. This rather "puristic" estimation strategy may reduce the precision of our estimates somewhat. It ensures, however, that the use of our data in empirical research will not create additional endogeneity problems, or mislead researchers to drawing tautological inferences. It will not "uncover" information that actually was used for estimating this data.

The data can be downloaded freely from http://hdl.handle.net/1902.1/15351. The downloadable dataset, available as Excel or ASCII files, comprises a balanced panel of ready-to-use annual data (population shares) on educational attainment by 3,076 U.S. counties (excluding Alaska and Hawaii) for 16 years (1990 – 2005). The data for 1990 and 2000 are from the decennial censuses,[2] and the data for 1991 – 1999 and 2001 – 2005 are estimated as described in this paper. The following Section 2 describes the estimation procedure, Section 3 discusses the reliability of our estimates, and Section 4 concludes.

2 Estimation

This section describes the method of estimating of the shares of residents in U.S. counties aged 25 or more by three educational groups for the intercensal years. The three education groups are residents holding a bachelor degree or higher, h_{rt}^{high}, residents holding a high-school diploma or higher but no bachelor degree, h_{rt}^{med}, and residents holding less than a high-school diploma, h_{rt}^{low}. We estimate these shares from three pieces of information: (i) the educational attainment of residents aged 25 or more by county in the census years, 1990 and 2000, (ii) the educational attainment (three groups) of residents aged 25 or more by state in the intercensal years, and (iii) total population aged 25 or more by county in the intercensal years. All these population data are available from the USCB.

[1] See http://factfinder.census.gov/servlet/DatasetMainPageServlet?_program=ACS&_submenuId=&_lang=en&_ts=.

[2] See http://factfinder.census.gov/servlet/DatasetMainPageServlet?_program=DEC&_submenuId=datasets_4&_lang=en&_ts=.

In terms of a county-by-education group matrix for each state and year, we have information on the row and column totals for the intercensal years but need to estimate the entries of the individual cells. The entries of the individual cells are known only for the two census years.

A variety of spatial disaggregation methods have been discussed in the literature to tackle estimation problems like this. See, for example, Li et al. (2007) or Wu et al. (2005) for recent surveys of such methods. Much emphasis has been put in this discussion on the question of how to efficiently use ancillary information on the geographical and other characteristics of the disaggregate spatial units. This question is of limited relevance in the present case because we deliberately use no ancillary information except total population by county. In addition to this, our estimation problem involves the time dimension in addition to the spatial dimension. We therefore use a rather simple two-step interpolation method. In the first step, we interpolate, separately for each state, the shares of residents in each skill group and county in total state population linearly over time. We assume that these shares change smoothly between the two census years and stay constant afterwards. The preliminary estimates we obtain from this linear interpolation over time do not meet the "pycnophylactic condition" (Tobler 1979). They typically sum up neither to total county population nor to total state population in the respective education group. We therefore use, in the second step, a simple nonlinear program that matches the preliminary estimates from the first step to total county populations and state-level shares of skills groups in total populations.

Formally, let M_{rsjt} denote the number of residents in education group j ($j = high, med, low$) in county r ($r = 1, ..., N_s$, N_s: number of counties in state s) of state s ($s = 1, ..., 49$) at time t, $M_{st} := \Sigma_{r=1}^{N_s} \Sigma_j M_{rsjt}$ total state population, and $\eta_{rsjt} := M_{rsjt}/M_{st}$ the share of residents in skill group j and county r in total state population. Using $\eta_{jrs1990}$ and $\eta_{rsj2000}$ available from the censuses, we linearly interpolate the preliminary county-skill group shares η_{jrst} in the first step by

$$\eta_{rsjt}^{prel} = \begin{cases} \eta_{rsj1990} + \left(\eta_{rsj2000} - \eta_{rsj1990}\right) \cdot \frac{t-1990}{10} & \text{for } t = 1991, ..., 1999 \\ \eta_{rsj2000} & \text{for } t = 2001, ..., 2005. \end{cases} \tag{1}$$

Since we do not have information on changes of the population shares after 2000, we simply assume them to be constant. The shares η_{rsjt}^{prel} are our preliminary estimates of the annual shares of each county and education group in total state population. The corresponding absolute population numbers, $M_{rsjt}^{prel} = \eta_{rsjt}^{prel} M_{st}$, sum up across all skill groups and counties to total state population in each year by construction. They do, however, not necessarily sum up across counties to the state-level population numbers by skill group, or across skill groups to total county populations.

To meet the pycnophylactic condition, we employ in the second step a simple nonlinear program for each state and intercensal year, which can be written as

$$min\ F = \sum_{r=1}^{N_s} \sum_{j=1}^{3} \eta_{rsjt}^{prel} (X_{rsjt} - 1)^2 \tag{2}$$

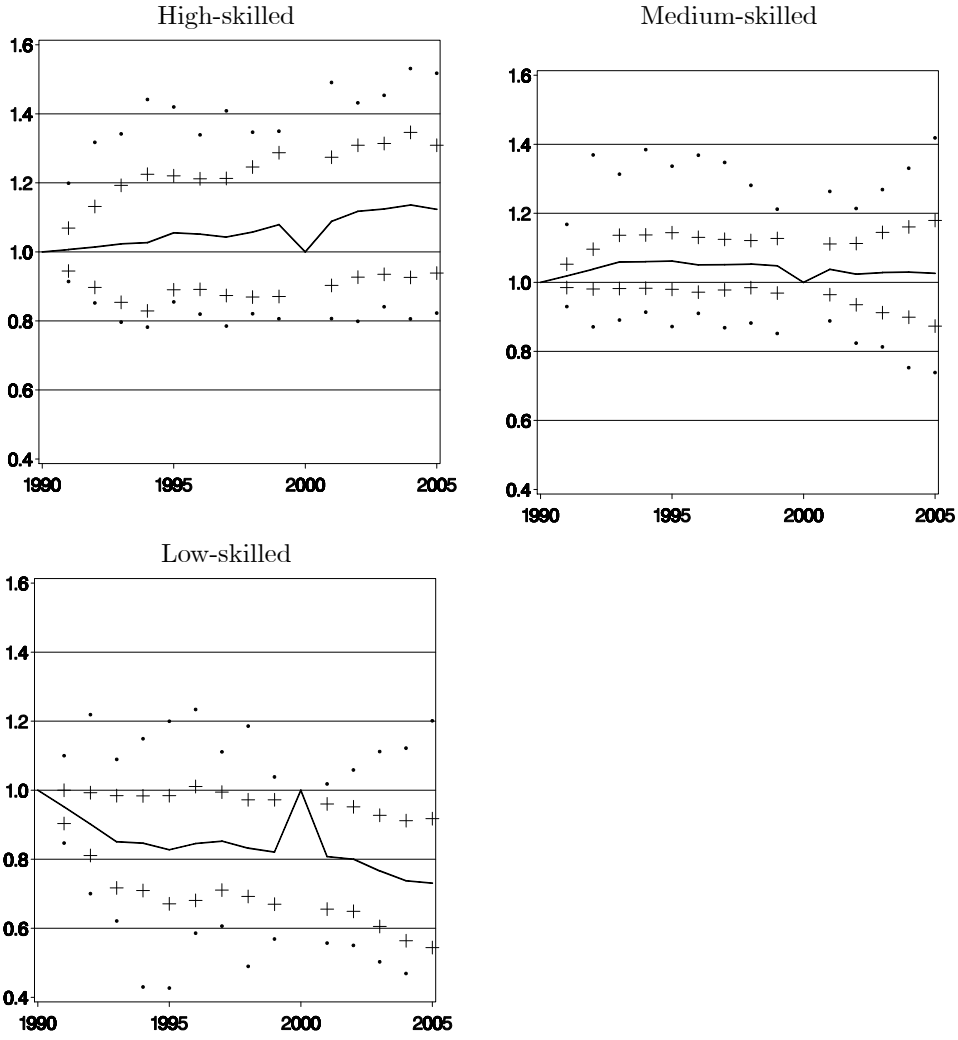
$$\begin{aligned} s.t.\ \sum_{r=1}^{N_s} \eta_{rsjt}^{prel} X_{rsjt} &= M_{jst} \\ \sum_{j=1}^{3} \eta_{rsjt}^{prel} X_{rsjt} &= M_{rst} \\ X_{rsjt} &\geq 0 \end{aligned}$$

$s = 1, ..., 49$; $t = 1991-1999, 2001-2005$.[3] M_{jst} denotes total population in skill group j in state s, and M_{rst} total population in county r within state s. Notice that this program does not impose any restrictions related to the autocorrelation of educational attainment over time or across counties. This program yields, for each state and year, an $(N_s \times 3)$ matrix of adjustment parameters $\widehat{X}_{rsjt}$ from which we calculate our final estimates of shares of the skill groups in total county population as $h_{rsjt}^{fin} = \eta_{rsjt}^{prel} \widehat{X}_{rsjt} M_{st} / M_{rst},\ j = high, med, low$. The corresponding absolute population numbers, $M_{rsjt}^{fin} = h_{rsjt}^{fin} M_{rst}$, match both the observed skill group totals by state and the observed county population totals in each year while differing as little as possible from the distribution of the corresponding preliminary estimates, M_{rsjt}^{prel}.

Figure 1 plots selected descriptive statistics for the adjustment parameters $\widehat{X}_{rsjt}$ separately for each skill group: the annual means, 95% confidence intervals around these means, and minima and maxima. For expositional convenience, the values of $\widehat{X}_{rsjt}$ are set to one for the census years where no estimation is needed. The figure shows that these adjustment parameters are, except for a few extreme values, generally fairly close to one for the high-skilled (bachelor degree or higher) and the medium-skilled population (high-school diploma, no bachelor degree). For the low-skilled population, they are mostly below one, especially for later years. The size of this group may be somewhat underestimated, which implies that the sizes of the higher-skilled groups may be somewhat overestimated.

[3] The USCB's total state population numbers differ slightly between the two statistics of state-level skill group and county-level population estimates. We assume that the state-level population estimates are more accurate than the county-level estimates. The total county populations are therefore determined by multiplying the share of each county in total state population, calculated from the USCB county-level estimates, by total state population, as given by the state-level estimates. State-level educational attainment data for the three skill groups used here is not available for the years 1991 and 1992. They are estimated by linear interpolation from the corresponding data for the years 1990 and 1993.

Figure 1: Descriptive statistics for the estimated adjustment parameters $\widehat{X}_{rsjt}$.



Notes: The solid lines denote the means, "+" the 95% confidence intervals, and "·" the maxima and minima for each year.

3 Plausibility checks

We check the plausibility of our estimates in three ways. First, we check how closely the estimated educational attainment data for the intercensal years are correlated across counties with the known educational attainment data for the

census years. While these correlations can be expected to decrease with growing time span between an intercensal year and a census year, they should decrease rather smoothly. Strong fluctuations of the correlation coefficients over time will raise doubts about the reliability of our estimates. Recall that the non-linear program (see equation 2) that fits the county-skill-level estimates to the county and state totals does not include any restrictions that account for the autocorrelation of educational attainment over time or across counties. Figure 2 plots, separately for each education group, the Pearson correlation coefficients between our annual estimates and the data for the two census years. The solid (dotted) lines represent the time series of correlation coefficients between our annual estimates and the education attainment data in 1990 (2000). The figure shows that the correlations between our estimates and the known data are very high and evolve rather smoothly over time for all three education groups. For example, the correlation with the 1990 census data (solid lines) decreases almost continuously over time toward the correlation coefficient between the 1990 and 2000 census data, which is about 0.9 for high- and low-skilled population and about 0.85 for medium-skilled population. The only notable outlying year is 1999, where our estimates correlate somewhat less with the known data from the two census years for all three education groups.

Second, we check if our estimates of educational attainment yield plausible regression results. We estimate a simple regional wage equation that can be derived from the human-capital augmented regional production function

$$Y_r = A_r \left(h_r^{\gamma} L_r\right)^{\alpha} K_r^{\beta} \tag{3}$$

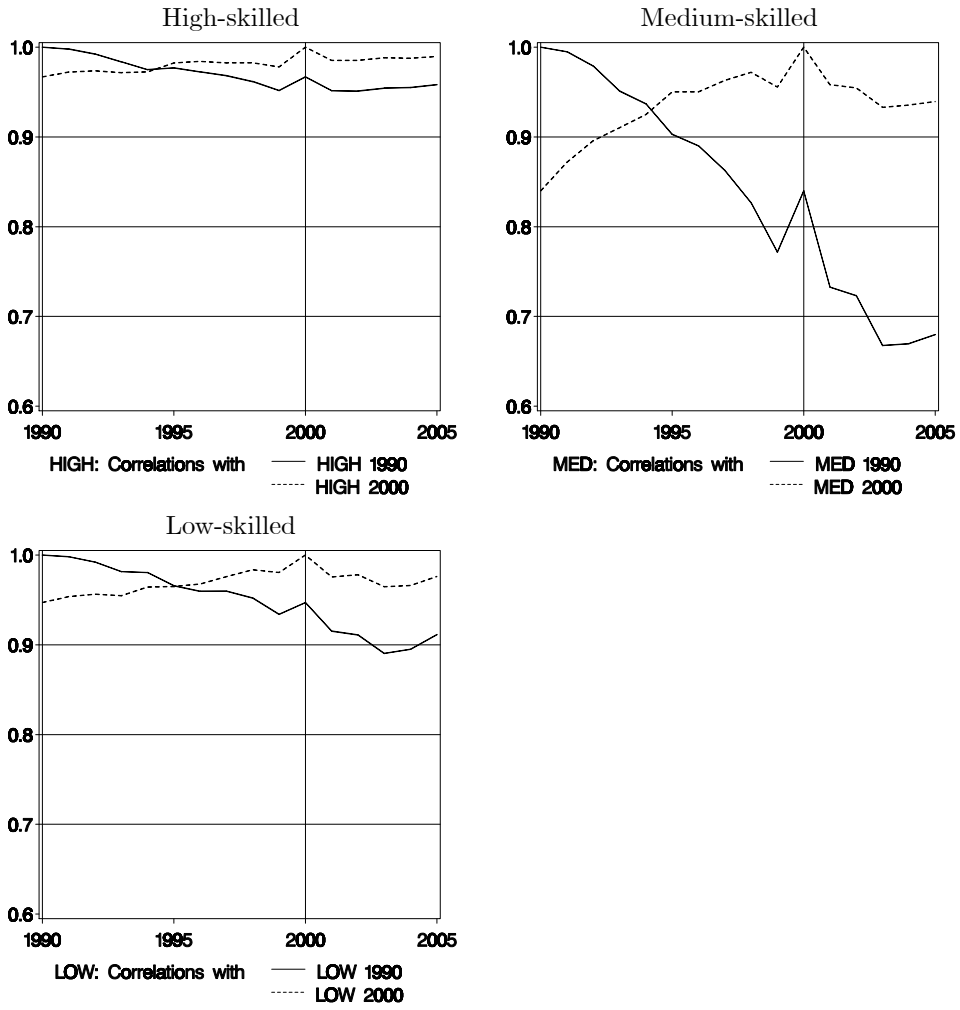
where Y_r, L_r and K_r denote regional output, regional labor input and regional physical capital input. A is total factor productivity, which we assume to vary only randomly across regions for simplicity (i.e., $A_r = Ae^{\varepsilon_r}$), and h_r is the human-capital intensity of the regional workforce, which we proxy by our estimated educational attainment shares. We eliminate physical capital from (3) by using its first-order condition, $r = \beta Y_r / K_r$, assuming the rental rate of capital, r, to be equalized across all regions by capital mobility. We then use the first-order condition $\partial Y_r / \partial L_r = w_r$ to obtain our log-linear regression equation

$$\ln w_r = c + \frac{\gamma\alpha}{1-\beta} \ln h_r + \frac{\alpha+\beta-1}{1-\beta} \ln L_r + \varepsilon_r. \tag{4}$$

c is a constant term, and ε_r is an error term, which we assume to be i.i.d. and normally distributed. We estimate (4) year by year for a narrow and a wider definition of human-capital intensity, h_r. In the narrow definition, we measure it by the share of residents with bachelor degree or higher, our variable h_{rt}^{high} (see Section 2). In the wider definition, we measure it by the share of residents with high school degree or higher, $h_{rt}^{high} + h_{rt}^{med}$.[4]

[4] The wage rate, w_{rt}, is measured as (nominal) wage and salary disbursements divided by wage and salary employment (number of jobs). The data is from the Regional Economic Information System (REIS, Table CA34) of the Bureau of Economic Analysis (BEA, see

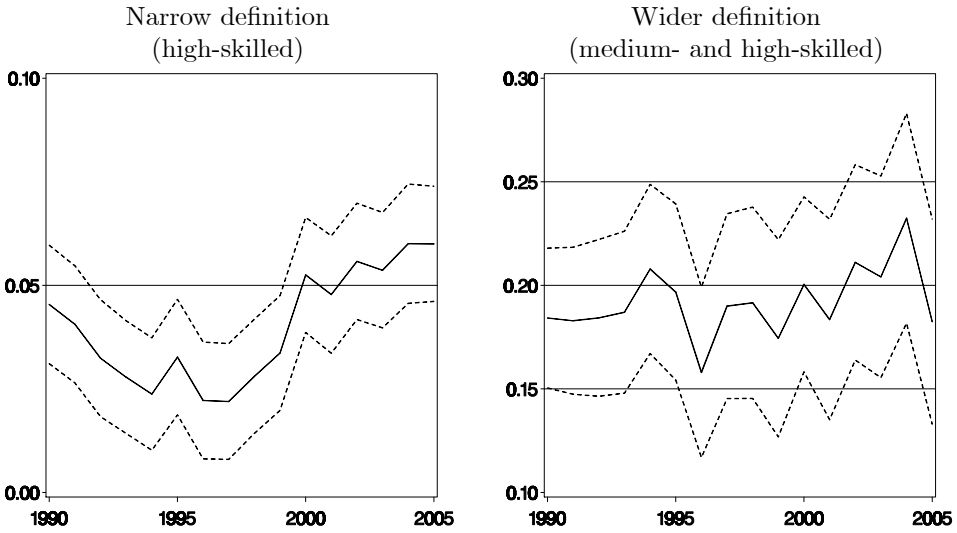
Figure 2: Correlations between estimated and known educational attainment data.



Note: Pearson correlation coefficients across 3,076 U.S. counties for education group j ($j = high, med, low$) between year t ($t = 1990 - 2005$) and the two census years (1990, 2000).

http://www.bea.gov/regional/index.htm). The data on employment, L_r, is also from the REIS database.

Figure 3: Parameters of human-capital intensity estimated from annual cross-section regressions.



Notes: Annual cross-section OLS estimations of (4) across 3,076 U.S. counties. The graphs depict the values (solid lines) and 95% confidence intervals (dotted lines) of estimated parameter of $\ln h_r$ in (4).

Figure 3 plots the annual parameter estimates and their 95% confidence intervals for these two definitions of human-capital intensity. The parameter of human capital in the narrow definition, depicted in the left panel of Figure 3, is estimated to be positive and significant for all years. The estimates for the intercensal years are somewhat lower during the 1990s, and somewhat higher during the 2000s than those for the census years 1990 and 2000, though. The parameter of human capital in the wider definition, depicted in the right panel of Figure 3, is also estimated to be positive and significant for all years and differs only little between the census and the intercensal years. This indicates that our estimates of educational attainment do a reasonable job in controlling for human-capital intensities in multiple regressions.

And third, we compare our estimates for the years after 2000 to the estimates published by the USCB for selected counties. In addition to the point estimates of the number of residents (aged 25 or higher) by several education groups, the USCB also publishes error margins for each estimate. We determine, separately for each year between 2001 and 2005 and for each of our three educational groups, the shares of counties for which our estimates lie outside the error margins published by the USCB. Table 1 reports the results. The first column (N) reports the number of counties for which USCB estimates are available. With the exception of the 2001 estimates, our estimates match those

of the USCB reasonably well. The shares of counties where our estimates don't match those of the USCB are below 10% in all cases and below 5% in most cases. Only in 2001, our estimates lie outside the USCB error margins in up to one third of the 21 counties. The reasons for this strong mismatch are subject to speculation. Maybe it is the USCB estimates rather than ours that is less reliable for this year.

Table 1: Mismatch shares between our and USCB estimates 2001 – 2005.

year	N	low-skilled	medium-skilled	high-skilled
2001	21	0.048	0.238	0.333
2002	232	0.005	0.022	0.073
2003	234	0.000	0.000	0.047
2004	237	0.000	0.025	0.063
2005	742	0.018	0.039	0.082

Notes: Shares of the "N" counties for which our estimates lie outside the confidence intervals given by the USCB.

4 Conclusions

This paper documents the method of estimating annual data on educational attainment of residents aged 25 or more in 3,076 (mainland) U.S. counties during the period 1991 – 2005. This data is designed specifically for use in panel regressions. It is purposefully estimated without resorting to ancillary information. This helps prevent the inferences drawn from the use of this data from being spurious or even tautological. The paper also reports a series of plausibility checks, which indicate that the estimates are correlated fairly highly with the corresponding data published by the USCB and yield plausible parameter estimates in regressions.

References

Li, T., D. Pullar, J. Corcoran and R. Stimson (2007), A Comparison of Spatial Disaggregation Techniques as Applied to Population Estimation for South East Queensland (SEQ), Australia. Applied GIS 3 (9): 1–16.

Wu, S.-S., X. Qiu and L. Wang (2005), Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience and Remote Sensing*, 42: 58–74.