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

Eckhardt Bode*

April 19, 2011

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 use in panel regression analyses. Several plausibility checks indicate that the data is fairly reliable and yields plausible parameter estimates in regressions.

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 time-varying economic indicators, annual data is, however, available only for 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 fills in such a gap by estimating 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, 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

*Kiel Institute for the World Economy, Hindenburgufer 66, 24105 Kiel, Germany. Email: eckhardt.bode@ifw-kiel.de, Phone: +49(431)8814462, Fax: +49(431)8814500. The author would like to thank two anonymous referees for helpful comments and suggestions.

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

the USCB publishes 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—educational attainment data for the intercensal years 1991 to 1999 and 2001 to 2005, this study significantly extends the time dimension available for county-level panel regression analyses. To maximize the scope of regression analyses for which this data can be used, the paper explicitly refrains from using ancillary information in the estimations. It estimates educational attainment for the intercensal years only from the available data on educational attainment and the corresponding population totals. This rather "puristic" estimation strategy may reduce the precision of the estimates somewhat. It ensures, however, that the use of the estimates in regressions will not create additional endogeneity problems, or mislead researchers to drawing tautological inferences. The regressions will not "uncover" information that actually was used for estimating this data.

The data can be downloaded from http://hdl.handle.net/1902.1/15351 as Excel or ASCII files. It comprises a balanced panel for 3,076 U.S. counties (excluding Alaska and Hawaii) and 16 years (1990 – 2005) of ready-to-use annual shares of three mutually exclusive and exhaustive skill groups in total county population aged 25 or more: (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 high-school diploma. 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 the estimates, and Section 4 concludes.

2 Estimation

We estimate the shares of the three skill groups in total county population aged 25 or more for the intercensal years from three pieces of information: (i) educational attainment of residents aged 25 or more by county in the census years 1990 and 2000, (ii) educational attainment 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. In terms of a county-by-skill group matrix for each state and year, we know the full matrices for the census years but only the row and column totals for the intercensal years. The row totals are total county populations, the column totals populations by skill group in the state as a whole. We consequently need to estimate the entries of the individual cells of the matrices for all states and all intercensal years.

[2] We had to slightly adjust the original census data, which are available at http://factfinder.census.gov/servlet/DatasetMainPageServlet?_program=DEC& _submenuId=datasets_4& _lang=en& _ts=, to remove mismatches between these census data and data from other sources used in this study (see below for more detail).

Four aspects put this estimation problem apart from the standard estimation problems discussed in the geostatistics or related literature (e.g., Chiles and Delfiner 1999, Wackernagel 2003, Li et al. 2007, Wu et al. 2005). The first aspect is that this estimation problem requires interpolation over time but disaggregation over regions and skill groups.[3] There is a variety of approaches that interpolate data simultaneously over time and space (Wackernagel 2003). But we are not aware of a geostatistical approach that interpolates data over time and simultaneously disaggregates them over space (or skills). We therefore have to separate the time dimension from the two cross-section dimensions and focus on just one estimation method at a time, interpolation or disaggregation.

The second aspect is that even if we focused on disaggregation alone, we would still face a two-dimensional disaggregation problem. We would need to estimate, separately for each year, the distribution of the population across counties and skill groups simultaneously in a way that ensures that all relevant adding-up constraints ("pycnophylactic conditions"; Tobler 1979) are met. The estimated county-by-skill group populations should sum up across counties to the observed state-by-skill group populations, and also across skill groups to total county population. We are not aware of a disaggregation approach that facilitates simultaneous disaggregation of several interrelated variables. If we focused instead on interpolation alone, we would have a similar problem with the pycnophylactic conditions. We therefore have to adopt a two-step procedure where we estimate the population data in the first step by either disaggregation over space, disaggregation over skills, or interpolation over time, and then modify these estimates in a second step such that they meet the pycnophylactic conditions.

The third aspect is that we do not want to use external, ancillary information on the geographic or socioeconomic characteristics of the counties. Especially spatial disaggregation approaches depend heavily on ancillary information (Li et al. 2007). Ancillary information on a county's relative location, geographic size, degree of urbanization, per-capita income, productivity or industry composition will possibly help improve our estimates. It will, however, invalidate the use of these estimates in studies that seek to explain one of these geographic or socioeconomic characteristics. Time interpolation approaches may generally also involve the use of ancillary information. This is, however, not feasible in the present case because we lack degrees of freedom. We have only two observed data points in time, 1990 and 2000.

And the fourth aspect is that we have information from the census years on precisely those quantities that we are seeking to estimate for the intercensal

[3] Time interpolation methods use linear or nolinear spline curves estimated from the available data points in time to determine the values of the unknown data points in time (e.g., Wackernagel 2003). Spatial disaggregation methods divide the data observed for "source zones" (states in the present case) among the "target zones" (counties), for which the data is unavailable, by estimating a weighting scheme that takes the relevant characteristics of the zones into account as far as possible (e.g., Flowerdale et al. 1991, Mennis 2003, Langford 2006). Both groups of methods may use ancillary information. Spatial disaggregation methods may, for example, use geographic or socioeconomic characteristics of the source and target zones that are expected to affect the spatial distribution of the data to be disaggregated.

years. This information may generally be used in both spatial disaggregation and time interpolation approaches. In a spatial disaggregation approach, we may use the census data to estimate a model, possibly a cross-section regression model estimated separately for each skill group, that explains the spatial distribution of each skill group across counties as closely as possible, and then use the parameters of this model to predict the spatial distributions of the skill groups in the intercensal years. In a time interpolation approach, we may interpolate the county-by-skill group populations separately for each county and skill group over time, using the two census years as fixed points. Compared to spatial disaggregation, this interpolation approach may yield reasonable results even without using ancillary information on the variety of factors that affect the skill compositions of counties. The observed county-by-skill group populations from the census years that serve as fixed points actually reflect the influences from all these factors implicitly. By interpolating the county-by-skill group populations over time, we implicitly project all these influences to the intercensal years.

This is the main reason why we prefer using the time interpolation approach separately for each county and skill group in this study.[4] We interpolate the share of each skill group in each county in total state population rather than absolute population by county and skill group in order to make the estimates less sensitive to state-wide business cycle fluctuations or shocks.[5] We denote

[4] Following the suggestion by a referee, we also employed a spatial disaggregation approch as a robustness check. We limited the ancillary information taken into account in this approach to geographical information that allowed us estimating a spatial lag parameter. We estimated, separately for each census year and each state, a cross-section spatial lag model (see Anselin 1988) where the share of each county and skill group in total state population is explained by three variables: the share of the county in total state population, the share of the skill group in total state population, and a spatially lagged dependent variable, which is the average of the county-by-skill group population shares in the neighboring counties. We used first-order binary contiguity weights for the latter, which generally perform fairly well in detecting the true form of spatial dependence in regression models (Stakhovych and Bijmolt 2008). The results of this approach, which are not reported here because of space restrictions but are available from the author upon request, are clearly inferior to those of the preferred time interpolation approach. Even though the spatial lag model fits the data very well for most states and skill groups, many predicted county-by-skill group population shares even came out negative for the intercensal years. This is because the spatial lag parameter or the parameter of the state-by-skill group population share is estimated to be negative for several states. This rather disappointing result does, of course, not imply that the spatial disaggregation approach is generally inferior to the time interpolation approach. It shows, however, that spatial disaggregation needs much more ancillary information to generate satisfactory results.

[5] The three USCB data sets we use as inputs differ from each other in terms of the total state population numbers. For example, the annual county population estimates for 2000 differ from the census 2000 population estimates, and neither of these two population estimates sum up across counties to the values published by USCB as its 2000 state population estimates. Similarly, the census 2000 population estimates by skill group do not sum up across counties to the corresponding 2000 state population estimates. We therefore had to harmonize the population data before interpolating the county-by-skill group populations for the intercensal years. Assuming that the annual state-level estimates are more reliable than the county-level estimates, we used them as our reference. We adjusted all county-level data such that they sum up across counties to this reference. To make sure that this data also sum up across skill groups to total county population, we used a nonlinear program similar to that described in equation (2) below. State-level educational attainment data for the three skill groups used

this county-by-skill group share by $\eta_{rsjt} := M_{rsjt}/M_{st}$, where M_{rsjt} is the population in skill group j ($j = high, med, low$) and county r of state s in year t, and M_{st} ($= \Sigma_{r=1}^{N_s} \Sigma_j M_{rsjt}$) total population of state s ($s = 1, ..., 49$). N_s is the number of counties in state s. For the period 1991–1999, we assume the county-by-skill group shares η_{rsjt} to change smoothly along the linear trend between the two census years. For the period 2001–2005, by contrast, we assume them to be the same as in 2000.[6] Formally, we thus determine "preliminary" county-by-skill group population shares, η_{rsjt}^{prel}, by interpolation separately for each county and skill group as

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

where $\eta_{rsj1990}$ and $\eta_{rsj2000}$ are the corresponding observed shares from the two census years. The preliminary shares η_{rsjt}^{prel} sum up to one across all skill groups and counties in a state in each year by construction. They do, however, not necessarily sum up across counties to the state-by-skill group population share, or across skill groups to the county population share.

In the second step, we therefore use a nonlinear program to match these preliminary shares to the corresponding state-by skill group and county totals separately for each state and intercensal year. This program determines one multiplicative correction factor, denoted by X_{rsjt} ($X_{rsjt} \geq 0$), for each county-by-skill group population share. It uses a weighted least squares objective function that aims at keeping the adjustments necessary to match the estimates to the pycnophylactic conditions at a minimum. The pycnophylactic conditions enter this program as constraints. Formally, this program can—for state s and intercensal year t—be written as[7]

$$\underset{r,\ j}{min}\ F_{st} = \sum_{r=1}^{N_s} \sum_j \eta_{rsjt}^{prel} (X_{rsjt} - 1)^2 \tag{2}$$

$$\begin{aligned} s.t.\ \sum_{r=1}^{N_s} \eta_{rsjt}^{prel} X_{rsjt} &= \eta_{sjt}, && j = high, med, low; \\ \sum_j \eta_{rsjt}^{prel} X_{rsjt} &= \eta_{rst}, && r = 1, ..., N_s; \\ X_{rsjt} &\geq 0, && j = high, med, low, \quad r = 1, ..., N_s. \end{aligned}$$

here was not available for the years 1991 and 1992. We estimated them by linear interpolation from the corresponding data for the adjacent years 1990 and 1993.

[6] As a robustness check, we extended the trend from the 1990s to the 2000s rather than setting it to zero for the years after 2000. This modification does not affect the results to a notable extent.

[7] We use the NLP procedure in SAS to run the estimations. The code is available from the author upon request.

The first row of (2) gives the objective function, which minimizes the weighted sum (over all skill groups and counties in state s) of the squared deviations of the correction factors from one. With $\eta^{prel}_{rsjt} X_{rsjt} = \eta^{fin}_{rsjt}$ being our final estimates for the county-by-skill group population shares, the $3N_s$ estimated correction factors, $\widehat{X}_{rsjt}$, $j = high, med, low$, $r = 1, ..., N_s$, determine by which percentage each county-by-skill group population share, estimated in the first step, has to be adjusted to ensure that all $3N_s$ shares together meet the pycnophylactic conditions. The second and third row of (2) give the pycnophylactic conditions as constraints. Taken literally, there are $3 + N_s$ conditions, namely one for each skill group and one for each county. η_{sjt} denotes the share of skill group j in total population of state s, and η_{rst} the share of county r in total population of state s. The last row of (2), finally, gives the $3N_s$ non-negativity conditions. Having estimated all correction factors, we calculate the shares of the skill groups in total county population as $h^j_{rsjt} = \eta^{prel}_{rsjt} \widehat{X}_{rsjt} M_{st}/M_{rst}$.[8]

Figure 1 plots descriptive statistics for the adjustment parameters $\widehat{X}_{rsjt}$ across all 3,076 counties for all years and separately for each skill group. It plots the annual means, 95% confidence intervals around these means, and minima and maxima. The values of $\widehat{X}_{rsjt}$ are one, and their variances zero, for the census years for which 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; Figure 1a) and the medium-skilled populations (high-school diploma, no bachelor degree; Figure 1b). For the low-skilled population (Figure 1c), the adjustment parameters vary somewhat more and are mostly below one for several years in the mid 1990s and mid-2000s. This indicates that the nonlinear program allocates more mass from this skill group to other skill groups in these years in order to match the first-step estimates, η^{prel}_{rsjt}, to the corresponding state and county totals.

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 observed educational attainment data for the census years. While these correlations can be expected to decrease with increasing time span between an intercensal year and a census year because of regionally differing structural changes in the skill compositions of the populations, they should decrease smoothly over time. The distribution of a skill group across counties estimated for, say, 1993 should be more similar to that in the census year 1990 than the distribution estimated for 1997, and that estimated for 1997 should be more similar than that observed for the census year 2000. Strong fluctuations of the correlation coefficients over time will, by contrast, raise doubts about the reliability of our estimates. They may indicate that the

[8] The downloadable Excel or ASCII files also report our estimates of total county population (see footnote 5) in order to facilitate calculation of absolute population figures by skill group.

nonlinear program employed in the second stage is subject to some sort of a "multiple equilibria" problem. Since it is not restricted to account for autocorrelation over time or across counties, it may converge to distributions that differ fundamentally from one year to another. Figure 2 plots, separately for each education group, the Pearson correlation coefficients between our annual estimates and the observed data for the two census years. The solid (dotted) lines represent the time series of correlation coefficients between our annual estimates and the observed education attainment data in 1990 (2000). The figure shows that the correlations between our estimates and the observed census data evolve rather smoothly over time for all three education groups. The correlation coefficients with the 1990 census data (solid lines), for example, decrease almost continuously over time toward the correlation coefficients between the two census years 1990 and 2000, which are about 0.95 for high-skilled (Figure 2a), 0.92 for low-skilled (Figure 2c) and about 0.75 for medium-skilled population (Figure 2b). This indicate that multiple equilibria are not a problem of the nonlinear program employed in the second stage of our estimation.

Second, we check if our estimates of educational attainment yield plausible regression results. For this purpose, we estimate a regional wage equation derived from a human-capital augmented regional production function separately for each year between 1990 and 2005 and check how strongly the parameters of human capital estimated for the intercensal years deviate from those estimated for the census years. Assuming that the true output elasticities of human capital did not change over time, the estimates for the intercensal years should not deviate too much from those for the two census years. We derive the regional wage equation to be estimated from the 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 output, labor input and physical capital input in county r. A is total factor productivity, which we assume to vary only randomly across regions for simplicity, and h_r is the human-capital intensity of the regional workforce, which we proxy, for the census years, by the observed educational attainment shares, or, for the intercensal years, by our estimated 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 counties 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 normally distributed with zero mean and constant variance. We estimate (4) for the cross section of all 3,076 counties year by year for a narrow and a wider definition of human-capital intensity, h_r. For the narrow definition, we measure human capital intensity by the share of residents with bachelor degree or higher, our variable h_{rt}^{high} (see Section 2). For the wider definition, we measure it by the share of residents with high school degree or higher, $h_{rt}^{high} + h_{rt}^{med}$.[9]

Figure 3 plots the annual parameter estimates and the bounds of their 95% confidence intervals for these two definitions of human-capital intensity. The parameters of human capital in the narrow definition, depicted in panel a of Figure 3, are estimated to be positive and significantly different from zero for all years. The estimates for the intercensal years in the 1990s are, however, rather low. In several of these years, the point estimates are actually below the lower bounds of the 95% confidence intervals of the corresponding estimates for the census years 1990 and 2000. This indicates that our estimates for the population with bachelor degree or higher during the 1990s may lack precision to some extent. Since their parameters are still significantly different from zero, they nonetheless capture at least some of the wage effects of human capital. The parameters of human capital in the wider definition, depicted in panel b of Figure 3, are also estimated to be positive and significant for all years and differ only little between the census and the intercensal years. This indicates that our estimates of medium- and high-skilled population taken together capture the wage effects of human capital to a similar extent as the official data from the census years.

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 skill 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 rather strong mismatch in 2001 are subject to speculation. Maybe it is the USCB estimates rather than ours that is less reliable for this year. As a further illustration, Figure 4 maps, for the high-skilled population in 2005, the 742 counties where USCB estimates are available. The 681 counties where our estimates do not differ significantly from the USCB estimates (i.e., lie within the error margins) are in light grey, the 5 where our estimates are significantly below the USCB estimates in medium grey, and the 56 where our estimates are significantly above the USCB estimates in black. The figure shows, first, that the USCB estimates are concentrated on the west coast including Arizona and New Mexico, New England, the Lake Michigan area and Florida. And second, it shows that those counties where our estimates do

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

not match those by the USCB are not strongly clustered in space.

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 suited particularly well for use in panel regressions. It is purposefully estimated without resorting to ancillary information on the geographic or socioeconomic characteristics of the counties. The use of ancillary information may cause additional endogeneity problems in the regressions using this data, and may tempt users into drawing tautological inferences. Even though it goes without ancillary information, the method yields fairly plausible estimates, as a series of plausibility checks suggest. The checks indicate, among others, that the estimates are correlated fairly highly with similar data published by the USCB for selected years and counties, and that they yield similar parameter estimates in regressions as the observed census data.

References

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Chiles, J.-P., and P. Delfiner (1999), *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Flowerdale, R., M. Green and E. Kehris (1991), Using Areal Interpolation Methods in Geographic Information Systems. *Papers in Regional Science* 70 (3): 303–315.

Langford, M. (2006), Obtaining Population Estimates in Non-census Reporting Zones: An Evaluation of the 3-class Dasymetric Method. *Computers, Environment and Urban Systems* 30 (2): 161–180.

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.

Mennis, J. (2003), Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer* 55 (1): 31–42.

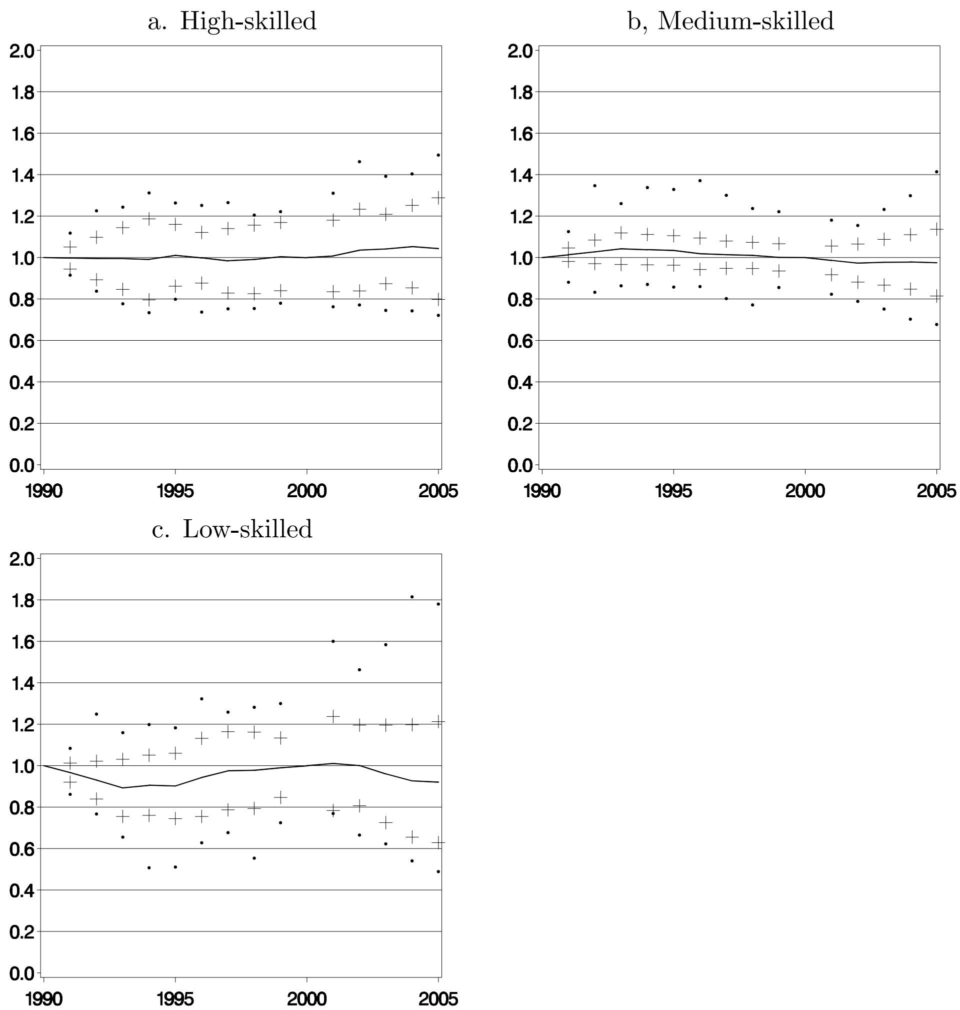
Stakhovych, S., and T.H.A. Bijmolt (2008), Specification of Spatial Models: A Simulation Study on Weights Matrices. *Papers in Regional Science* 88 (2): 389–408.

Tobler, W.R. (1979), Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association* 74 (367): 519–530.

Wackernagel, H. (2003), *Multivariate Geostatistics*. 3rd edition. Berlin: Springer.

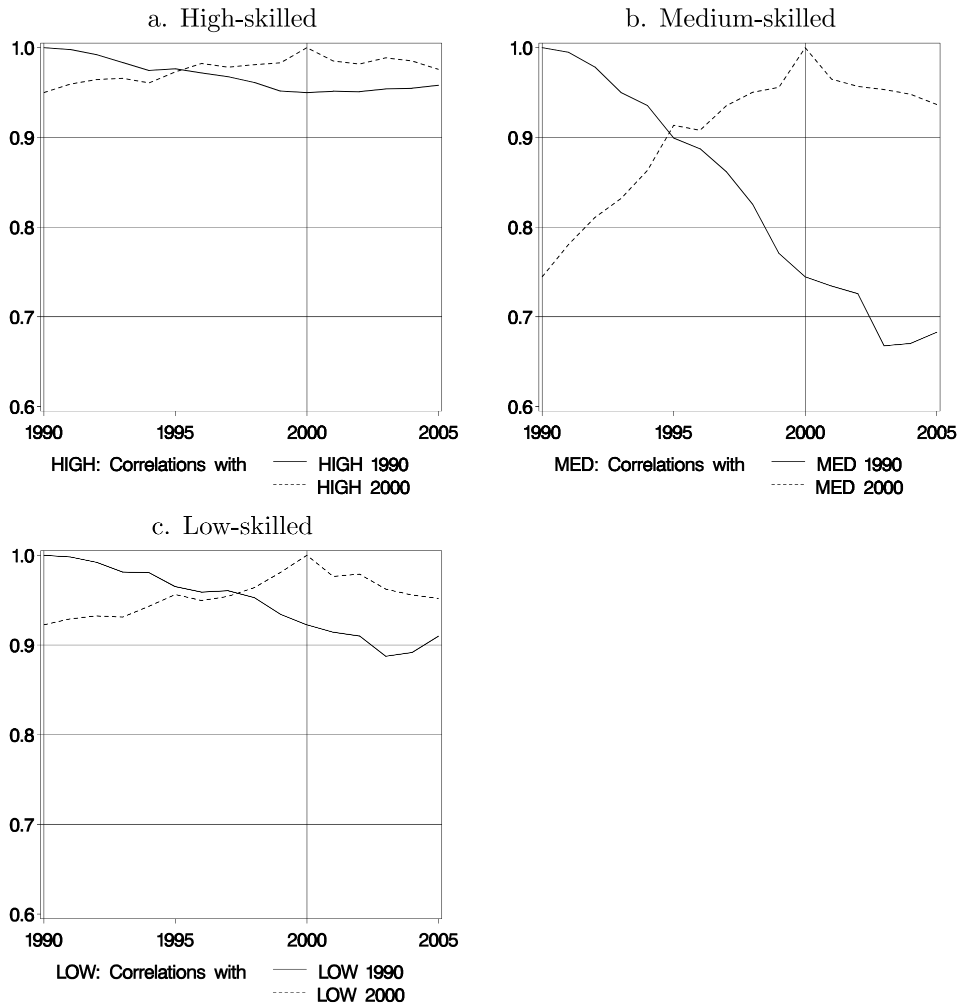
Wu, S.-S., X. Qiu and L. Wang (2005), Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience and Remote Sensing*, 42 (1): 58–74.

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



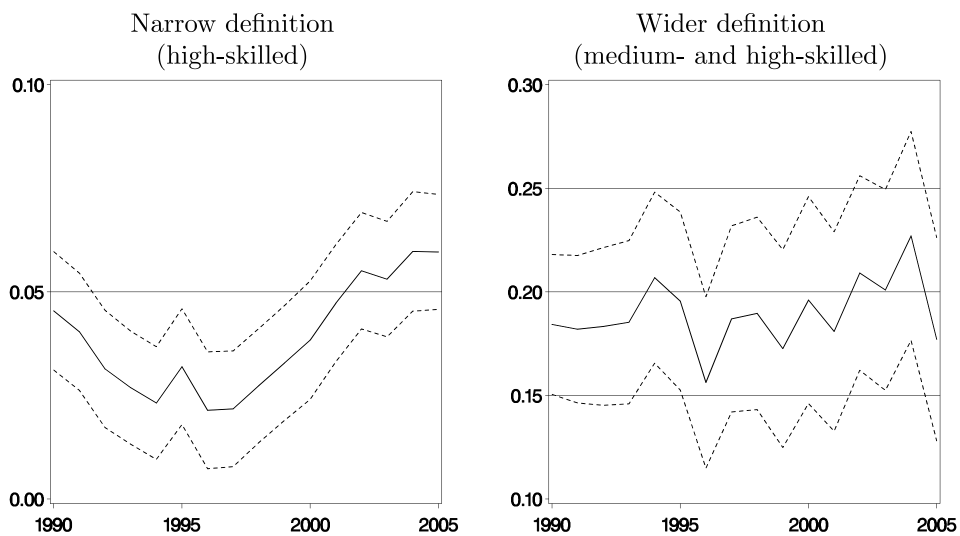
Notes: The solid lines denote the annual means, "+" the bounds of the corresponding 95% confidence intervals, and "·" the maxima and minima of the adjustment parameters for the respective skill groups across the 3,076 counties.

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



Note: Solid lines (dotted lines): Pearson correlation coefficients across 3,076 U.S. counties between the share of the respective skill group in total county population in census year 1990 (2000) and the corresponding estimated share in the year depicted on the horizontal axis.

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



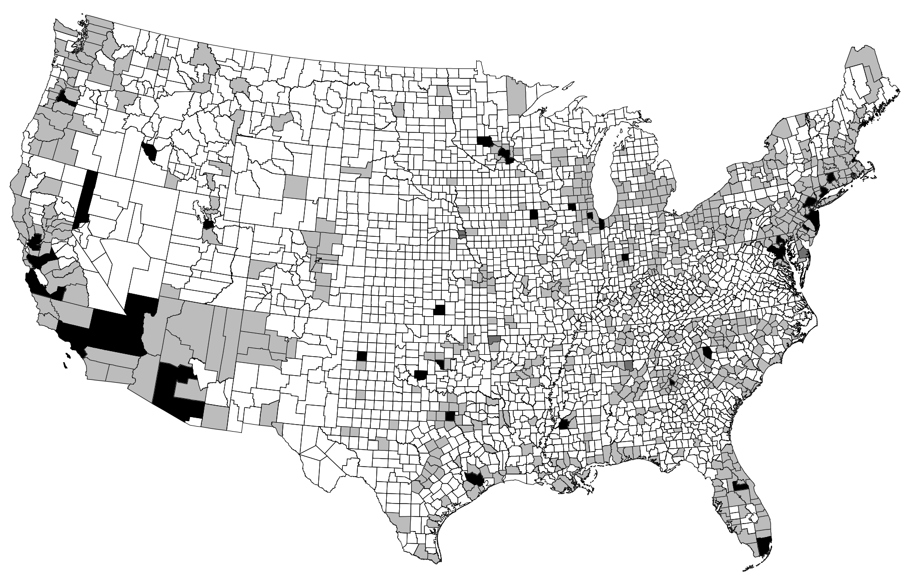
Notes: Annual cross-section OLS estimations of (4) across 3,076 U.S. counties. The graphs depict the values (solid lines) and the bounds of the 95% confidence intervals (dotted lines) of the annual estimates for the parameter of $\ln h_r$ in (4).

Table 1: Mismatches 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	0.004	0.022	0.073
2003	234	0.000	0.000	0.047
2004	237	0.000	0.025	0.063
2005	742	0.016	0.039	0.082

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

Figure 4: Mismatches between our and USCB estimates for high-skilled population 2005.



Notes: Light grey: Counties where USCB estimates are available, medium grey (black): counties where our estimates lie below (above) the error margins given by the USCB.