JOURNAL OF REGIONAL SCIENCE

JOURNAL OF REGIONAL SCIENCE, VOL. 56, NO. 5, 2016, pp. 868–894

COUNTERFACTUAL SPATIAL DISTRIBUTIONS*

Paul E. Carrillo

Department of Economics, The George Washington University, 2115 G St. NW Suite 364, Washington, DC 20052; and Elliot School of International Affairs, The George Washington University. E-mail: pcarrill@gwu.edu

Jonathan L. Rothbaum

U.S. Census Bureau, Income Statistics Branch, Social, Economic, and Housing Statistics Division, 4600 Silver Hill Rd, Washington, DC 20233. E-mail: jonathan.l.rothbaum@census.gov

ABSTRACT. Recent contributions provide researchers with a useful toolbox to estimate counterfactual distributions of scalar random variables. These techniques have been widely applied in the literature. Typically, the dependent variable of interest has been a scalar and little consideration has been given to spatial factors. In this paper we propose a simple method to construct the counterfactual distribution of the location of a variable across space. We apply the spatial counterfactual technique to assess how much changes in individual characteristics of Hispanics in the Washington DC area account for changes in the distribution of their residential location choices.

1. INTRODUCTION

The influential contributions of DiNardo, Fortin, and Lemieux (1996), Firpo, Fortin, and Lemieux (2009), Machado and Mata (2005), and Donald, Green, and Paarsch (2000) provide researchers with a useful toolbox to estimate counterfactual distributions of a scalar random variable of interest. These techniques have been used recently in several studies in urban, labor, public, and development economics.[1] In all applications in the literature, the dependent variable of interest has been a scalar (for example, wages, home prices, or consumption expenditures) and little (if any) consideration has been given to spatial factors. In this paper, we propose a simple method to construct the counterfactual distribution of the location of a variable across space.

Counterfactual distributions (or counterfactual distributional statistics) are at the core of decomposition methods in economics.[2] In the classic Oaxaca–Blinder wage

*We thank Daniel McMillen, Gianfranco Piras, Edson Severnini, Anthony Yezer, and participants at the conferences of the Real Estate and Urban Economics Association, Spatial Econometrics Association, Urban Economics Association, GWU Urban Day, and the Richmond Fed Regional Workshop for helpful comments and discussions. We are grateful to the George Washington Institute for International Economic Policy for research support. This work is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on (statistical, methodological, technical, or operational) issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Received: January 2015; revised: October 2015; accepted: December 2015.

[1] For applications in labor, see Albrecht, Bjorklund, and Vroman (2003), Dustmann, Ludsteck, and Schonberg (2009), Lemieux, MacLeod, and Parent (2009), and Lemieux (2006), among many others. Applications in urban economics include McMillen (2008), Nicodemo and Raya (2012), Carrillo and Pope (2012), Cobb-Clark and Sinning (2011), Fesselmeyer, Le, and Seah (2012, 2013), and Carrillo and Yezer (2009). For applications in public and development economics see, for example, Bound, Lovenheim, and Turner (2009), Pallais (2009), Dolls, Fuest, and Peichl (2012), and O'Donnell et al. (2009).

[2] For a comprehensive review about decomposition methods in economics see Fortin, Lemieux, and Firpo (2011).

© 2016 Wiley Periodicals, Inc.

DOI: 10.1111/jors.12277

decomposition researchers often simulate the counterfactual mean: how the mean wage of a demographic group would look like if they experienced the returns of a counterfactual group. More recent papers estimate the counterfactual at every point of the distribution. For example, Albrecht et al. (2003) use quantile decomposition techniques to simulate the counterfactual *distribution* of female wages: the wage distribution of females if they had the same demographic characteristics (endowments) as males. In a recent application in urban economics, McMillen (2008) simulates the *distribution* of home prices in Chicago in 2005 assuming that home characteristics remain constant as in 1995 (among other counterfactual simulations). He finds that the shift in home prices between 1995 and 2005 is significantly larger at the right tail of the distribution and that these shifts cannot be attributed to changes in the structural characteristics nor the location of the housing stock. In our paper, we show how to simulate a counterfactual *spatial* distribution: the distribution of a variable across space assuming that other observable characteristics are those from a counterfactual group.

The concept and usefulness of *spatial* counterfactual distributions can be illustrated with an example. Suppose the spatial distribution of a variable of interest, say, the residence location choices of Hispanics, has changed over time. Researchers are interested in assessing the relationship between changes in the spatial distribution of individuals' residences and changes in their observable characteristics (such as age, race and gender, and education, for example). A counterfactual spatial distribution can be a useful tool to assess this relationship. For instance, if the true data generating process (DGP) were known, one could simulate the distribution of the population across the urban area in the latter period, assuming that the distribution of demographic characteristics remains constant, as in the initial period. By comparing the actual and counterfactual distributions researchers can compute what portion of the change in the spatial distribution between the first and second period can be accounted for by changes in the characteristics of the population. If the true DGP were known, this exercise would be a straightforward parametric calculation. The true DGP, however, is generally unknown.

To compute counterfactual spatial distributions, we extend the semi-parametric decomposition methods developed by DiNardo et al. (1996). DiNardo, Fortin, and Lemieux (DFL) estimate the counterfactual distribution of a scalar random variable (wages) by computing a weighted empirical probability density function. The weights are a function of the frequency that the covariates appear in the actual data relative to the frequency of covariates in the counterfactual. In our paper, the actual and counterfactual *spatial* distributions are estimated as follows. We let y_1 and y_2 denote the location coordinates on a plane of a variable of interest. A conventional kernel method can be used to estimate the joint probability density function of random vector $[y_1, y_2]$, i.e., the spatial distribution. The same reweighting mechanism suggested by DFL is then used to estimate a weighted joint pdf and find the *counterfactual* spatial distribution. Numerical simulations using a known DGP confirm that our method is able to replicate the counterfactual distribution remarkably well.[3]

[3] A growing literature in economics uses permutation algorithms to simulate counterfactual "spatial" distributions and develop sampling distributions for tests statistics. For example, Duranton and Overman (2005, 2008) use such an approach to evaluate the extent of industries' localization. Duranton and Overman simulate the distribution of *bilateral distances* between each pair of establishments in an industry (after conditioning on the true spatial distribution of manufacturing) under the hypothesis that their location was randomly assigned and compare it with the observed distribution. A test statistic is developed and allows them to assess if an industry is localized or not. Our approach is rather different. We simulate the spatial distribution of individuals (or firms) in period t_1, assuming that the observed covariates remain

© 2016 Wiley Periodicals, Inc.

We then *illustrate* the spatial counterfactual technique to understand how changes in individual characteristics of Hispanics have affected their location choices in the Washington DC area. Our application shows that counterfactual spatial distributions are straightforward to compute and can provide interesting *reduced form* insights about the determinants of location choices of individuals. The rest of the paper is structured as follows. The next section presents the econometric methodology. In the third section we present our empirical application. The last section concludes.

2. METHODS

Basic Notation

Consider the random vector $[\, y_1, y_2,\ X]$, where y_1 and y_2 denote the location coordinates of a variable of interest and X is a random vector of relevant covariates. Let $T\ = t_0$ and $T\ = t_1$ refer to two mutually exclusive periods we analyze and $f(y_1,\ y_2, X|T\ = t_j)$ denote the joint probability density function in each period, where $j = \{0, 1\}$. To make our exposition concrete, we are going to use the same example used in the introduction, where the researcher is interested in computing the distribution of the population across space at two points in time. In this setting, $Y\ =\ [y_1, y_2|T]$ would measure the latitude and longitude where an individual resides in each period, and covariate X captures any variables that explains people's location choices (such as age, education, income, and race, for example).

The spatial distribution of the population in period t_1 is equivalent to the marginal joint density of random vector $[y_1, y_2|T\ = t_1]$,

$$(1) \qquad f(y_1, y_2|T = t_1)\ = \int f(y_1, y_2|x, T = t_1)\, h\, (x|T = t_1)dx,$$

where $h(x|T\ = t_1)$ is the marginal distribution of random variable X in period 1. Notice that T is a random variable describing the period from which an observation is drawn and x is a particular draw of observed attributes of individual characteristics from random vector X. $f(y_1, y_2|x, T\ = t_1)$ is the spatial distribution of our variable of interest (population) given that a particular set of attributes x have been picked, and $h(x|T\ = t_1)$ is the probability density of individual attributes evaluated at x. The spatial distribution of the population in period t_0 is defined similarly

$$(2) \qquad f(y_1, y_2|T = t_0)\ = \int f\, (y_1,\ y_2|x, T = t_0)\, h\, (x|T = t_0)\, dx.$$

Because in both periods y_1 and y_2 are observed in the data, the spatial distributions can be easily estimated using any conventional parametric or nonparametric (kernel) method.

Suppose we would like to assess how the spatial distribution of the population in period t_1 would look like if the distribution of individual attributes X (for example, age, education, income, and race) were the same as in period t_0. This counterfactual spatial distribution is denoted as $f_{x^1 \to x^0}$ and expressed symbolically as[4]

$$(3) \qquad f_{x^1 \to x^0}\, (y_1, y_2)\ \ = \int f\, (y_1,\ y_2|x, T = t_1)\, h\, (x|T = t_0)\, dx.$$

constant as in period t_0. Potentially, both approaches could be combined and one could apply, say, the Duranton–Overman localization tests to one of our estimated counterfactual distributions.

[4] The subscript "$x^1 \to x^0$" indicates that the attributes data from period t_1 will be "replaced" by data from period t_0.

© 2016 Wiley Periodicals, Inc.

Parametric Example

The construction of the counterfactual spatial distribution in Equation (3) is straightforward if the DGP were known. To illustrate this, we specify a simple parametric model (described in more detail in Appendix A) of the location coordinates y_1 and y_2 based on a single characteristic X and "taste" parameters ϵ_1^j and ϵ_2^j. Formally, the location coordinates chosen by individual i in period j are given by $y_{1i}^j = \epsilon_{1i}^j\ X_i^j$ and $y_{2i}^j = \epsilon_{2i}^j\ X_i^j$, where time period $j = \{0, 1\}$. We assume that X is a random variable and that its distribution depends on a sole parameter σ_{μ_j} which measures the dispersion of the X covariate in each time period. We also let parameters ϵ_1^j and ϵ_2^j be random and their distribution depend only on parameter θ_j.

Given a set of parameter values and the assumption that $\sigma_{\mu_1} > \sigma_{\mu_0}$ and $\theta_1 > \theta_0$, we simulate 100,000 realizations of these random variables in each period and estimate the joint distribution of $[y_1, y_2]$

$$f\,(y_1, y_2|T\ = t_1) = \ f\big(y_1, y_2;\ \sigma_{\mu_1}, \theta_1\big)\,,$$

and

$$f\,(y_1, y_2|T\ = t_0) = \ f\big(y_1, y_2;\ \sigma_{\mu_0}, \theta_0\big)\,.$$

Results are shown in Figure 1 in panels A and B, respectively.[5] In the figures we have defined a set of "natural" boundaries delineated by a 100×100 grid. Within each square in the grid, we use a uniform kernel to determine the sample density. This is equivalent to calculating the average density in the square, much like would be the case if we mapped average population density by natural boundaries such as census block groups, tracts, or counties. We have chosen the uniform kernel method to compute the joint density for its simplicity and similarity to the case of geographic data in natural boundaries. In applying this technique, the choice of boundaries and two-dimensional kernels used are at the discretion of the researcher.

Note that the differences between these distributions can be partially explained by changes in the distribution of covariates $X\,(\sigma_{\mu_1} > \sigma_{\mu_0})$ and partially explained by changes in other location preferences $(\theta_1 > \theta_0)$. Counterfactual spatial distributions allow us to assess, for example, how the spatial distribution in period 1 would look if the distribution of covariates remained as in period 0. This can be computed as follows

$$f_{x^1 \to x^0}\ (y_1, y_2) = \ f\big(y_1, y_2; \sigma_{\mu_0}, \theta_1\big)\,. \tag{4}$$

Results are shown in the first panel of Figure 2. Panel A shows the "true" counterfactual with X drawn from the period 0 distribution $(\sigma_C = \sigma_0)$ and the location preference from period 1 $(\theta_C = \theta_1)$.

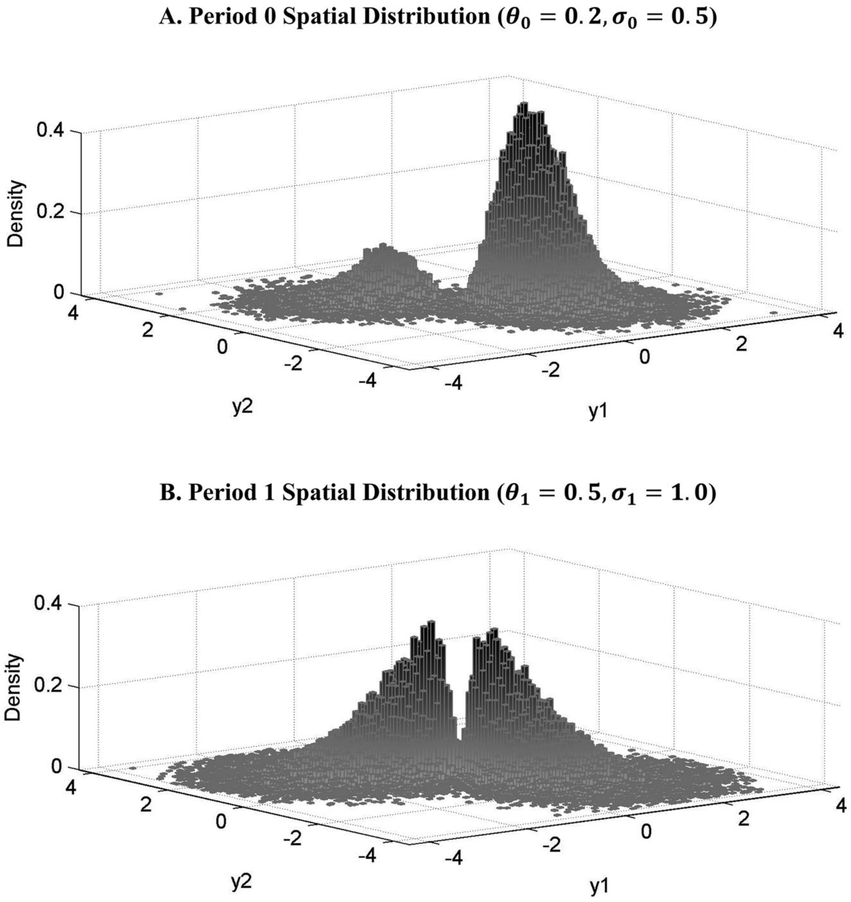
In practice, however, finding the "correct" DGP specification and the identification of structural parameters are very challenging tasks. For these reasons, parametric models are useful tools to understand the basic properties of spatial counterfactual distributions but are not often used in empirical applications.

Semi-Parametric Approach (DFL)

In order to compute the counterfactual distribution specified in Equation (3) without making strong parametric assumptions, we employ the semi-parametric decomposition

[5] The exact details of the parametric specification are discussed in Appendix A.

© 2016 Wiley Periodicals, Inc.



A. Period 0 Spatial Distribution ($\theta_0 = 0.2, \sigma_0 = 0.5$)

B. Period 1 Spatial Distribution ($\theta_1 = 0.5, \sigma_1 = 1.0$)

Notes: This figure shows the spatial distribution of individuals in our parametric example. Panel A shows the distribution in $T = t_0$ and Panel B shows the distribution in $T = t_1$. In period t_1 there is a change in the distribution of the observable X variable (determined by the parameter σ) and the unobservable "taste" parameter (determined by θ). In the figure, we have defined a set of "natural" boundaries delineated by a 100 × 100 grid. Within each square in the grid, we use a uniform kernel to determine the sample density. This is equivalent to calculating the average density in the square, much like would be the case if we mapped average density by natural boundaries such as census block groups, tracts, or counties.

FIGURE 1: Parametric Example.

methods developed by DFL. To keep our exposition self-contained, we provide a careful description of the DFL approach.

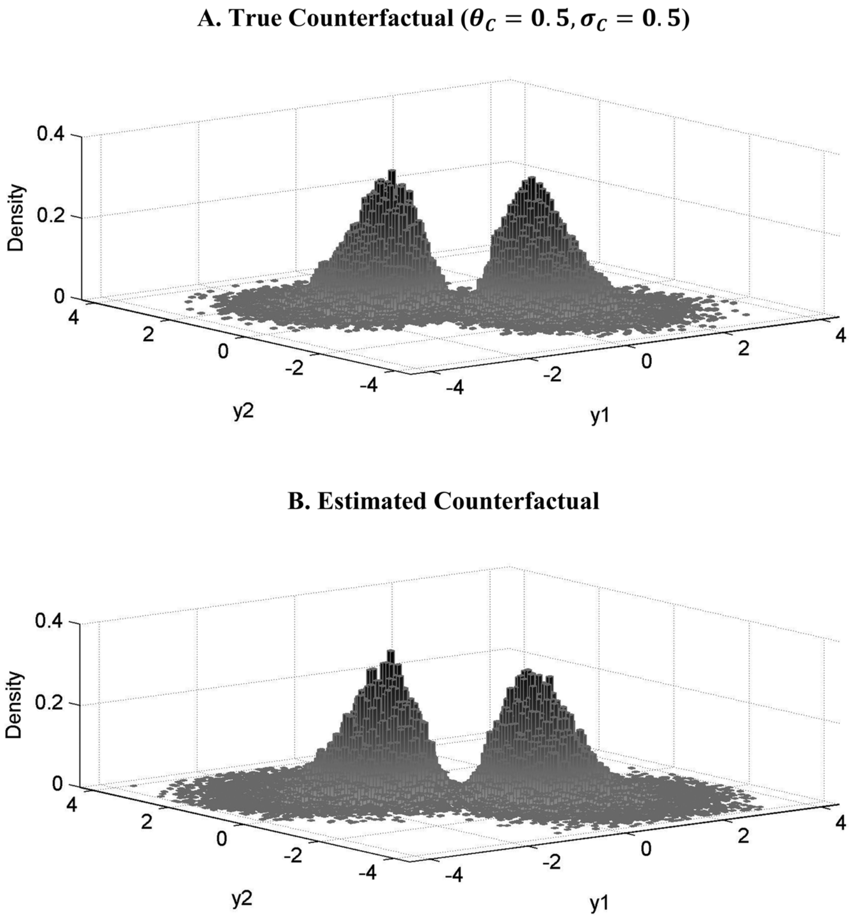
Bayes' rule is first used to obtain that $h\,(x|T\ =\ t) = \frac{P(X=x)P(T=t|X=x)}{P(T=t)}$. DFL recognized that

$$(5) \qquad \frac{h(x|T=t_0)}{h(x|T=t_1)} = \frac{\frac{P(T=t_0|X=x)}{P(T=t_0)}}{\frac{P(T=t_1|X=x)}{P(T=t_1)}} = \frac{\frac{P(T=t_0|X=x)}{1-P(T=t_0|X=x)}}{\frac{P(T=t_0)}{1-P(T=t_0)}} = \tau_{t_1\to t_0}(x).$$

One may use expression (5) to substitute $h(x|T = t_0)$ in Equation (3) and obtain that

$$(6) \qquad f_{x^1\to x^0}(y_1, y_2)\ = \int f(y_1,\ y_2|x, \mathrm{T} = \mathrm{t}_1)\,h\,(x|\mathrm{T} = \mathrm{t}_1)\,\tau_{t_1\to t_0}(x)\,dx.$$

© 2016 Wiley Periodicals, Inc.


A. True Counterfactual ($\theta_C = 0.5, \sigma_C = 0.5$)

B. Estimated Counterfactual

Notes: This figure shows the counterfactual spatial distribution of the individuals in our parametric example. Panel A shows the true counterfactual which uses the known data generating process with the unobservable "taste" parameter (determined by θ) from period t_1 and the observable characteristics X (determined by σ) from period t_0. In Panel B shows the counterfactual estimated using the DFL method. In the figure, we have defined a set of "natural" boundaries delineated by a 100×100 grid. Within each square in the grid, we use a uniform kernel to determine the sample density. This is equivalent to calculating the average density in the square, much like would be the case if we mapped average density by natural boundaries such as census block groups, tracts, or counties.

FIGURE 2: Counterfactual Spatial Distribution.

Notice that this expression differs from Equation (1) only by $\tau_{t_1 \to t_0}(x)$. DFL refer to $\tau_{t_1 \to t_0}(x)$ as "weights" that should be applied when computing the counterfactual distribution of our variable of interest. However, given that the weights are unknown, they need to be estimated.

To be specific, as in Carrillo and Pope (2012), we summarize the estimation algorithm for the counterfactual given that a random sample of N_0 and N_1 observations for periods t_0 and t_1 is available:

(1) Estimate $P(T = t_0)$ using the share of observations where $T = t_0$ to obtain $\hat{P}(T = t_0) = \frac{N_0}{N_0 + N_1}$.

© 2016 Wiley Periodicals, Inc.

(2) Estimate $P(T = t_0 \,|X = x)$ with a binary dependent variable model (such as probit or logistic regression) on all observations from both periods. The dependent variable equals one if $T = t_0$, and explanatory variables include the vector of individual attributes x.

(3) For the subsample of observations where $T = t_1$, estimate the predicted values from the binary dependent variable model. In the logit case, this is $\hat{P}(T = t_0 \,|X = x) = \frac{e^{x\hat{\beta}}}{1+e^{x\hat{\beta}}}$, where $\hat{\beta}$ is the parameter vector from the logit regression. Then, compute the estimated weights as

$$\hat{\tau}_{t_1\rightarrow t_0} = \frac{\frac{\hat{P}(T=t_0|X=x)}{1-\hat{P}(T=t_0|X=x)}}{\frac{\hat{P}(T=t_0)}{1-\hat{P}(T=t_0)}}.$$

(4) For the subsample of observations where $T = t_1$, compute the joint density of coordinates $[y_1, y_2|T = t_1]$ applying the estimated sample weights.

To illustrate the validity of this method we compute the counterfactual spatial distribution specified in Equation (6) using the DFL approach (that is assuming that the DGP is unknown). Results are shown in Panel B of Figure 2 using the same natural boundaries and uniform kernel as in the previous figures. The reweighting method is able to replicate the counterfactual spatial distribution remarkably well. To test for equality of distributions, we conduct a Pearson χ^2 goodness of fit test.[6] We are unable to reject the null hypothesis that the two distributions are the same.[7]

Decomposition

It may be useful to analyze the differences between the spatial distributions of interest. To assess how much of the changes in the joint distribution of $[y_1, y_2]$ between period t_0 and t_1 can be accounted for by changes in individual attributes x at a given point (y_1, y_2), we may compute

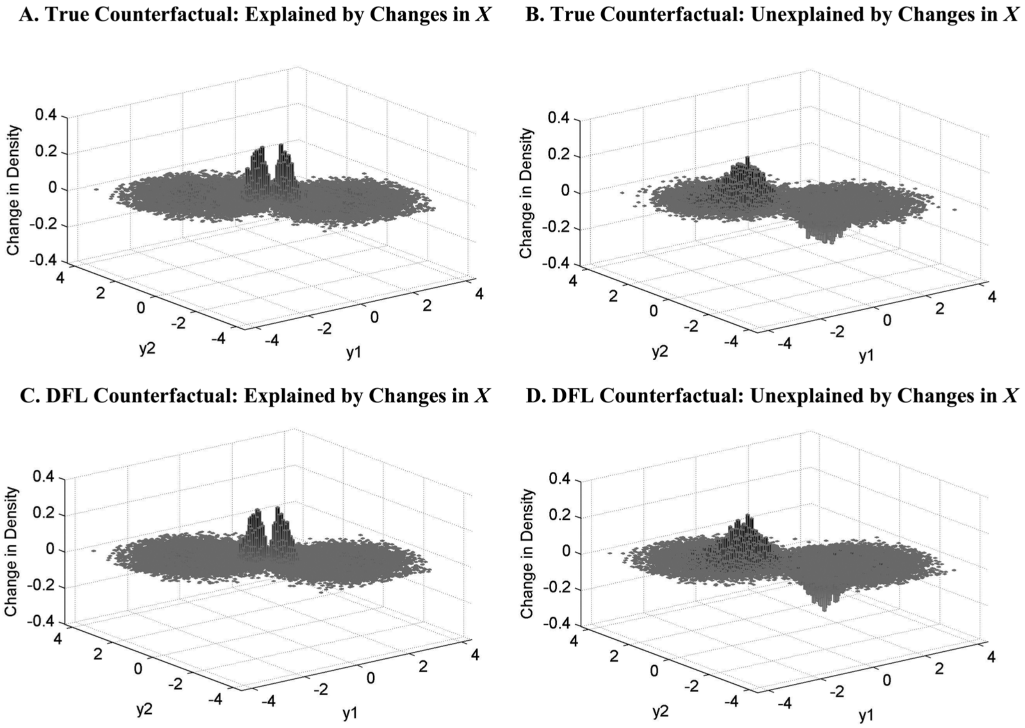
$$\begin{aligned} f(y_1,\ y_2|t_1) - f\,(y_1,\ y_2|t_0) = \ & \{f(y_1,\ y_2|t_1) - f_{x^1\rightarrow x^0}(y_1,\ y_2)\} \\ & - \{f(y_1,\ y_2|t_0) - f_{x^1\rightarrow x^0}(y_1,\ y_2)\}\,. \end{aligned} \tag{7}$$

The second term in brackets on the right-hand side of the equation above measures the "unexplained" part of the changes in the distribution of vector Y. Hence, the first term

[6] Using a 10×10 grid, we calculate a χ^2 statistic with a P-value of 0.30. The same test comparing the initial and final distributions to the true or DFL estimate of the counterfactuals rejects the null hypothesis that the distributions are the same with a P-value of virtually zero. Given the uniform kernel used in our application, a natural test for equality of distributions is the χ^2 test. However, alternative test statistics to compare two-dimensional empirical distributions are also available that do not require binning the data as in a histogram. Some examples include a two-dimensional Kolmogorov–Smirnov test (Fasano and Franceschini, 1987) and a test based on "statistical energy" (Aslan and Zech, 2005). The results for both of these tests are the same as for the corresponding χ^2 test for each set of distributions compared (only 1,000 observations are drawn in the samples for these tests as they are more computationally demanding than the χ^2 test).

[7] Changing the parameters of our toy model would affect the shape of the spatial distribution, but should not affect the ability of the DFL method to replicate the counterfactual distribution. We test this statement by allowing $\sigma_{\mu 2}$ and θ_1 to take any value in the intervals [0.75, 1.25] and [0.4, 0.6], respectively. We then randomly pick the values of σ_{μ_1} and θ_1, and test if the actual counterfactual distribution and the DFL estimated distribution are the same. In 1,000 simulations (with $n = 10{,}000$ and a 10×10 grid), we cannot reject the null that the distributions are the same almost every time.

© 2016 Wiley Periodicals, Inc.



A. True Counterfactual: Explained by Changes in X

B. True Counterfactual: Unexplained by Changes in X

C. DFL Counterfactual: Explained by Changes in X

D. DFL Counterfactual: Unexplained by Changes in X

Notes: This figure shows the decomposition of the change in the distribution between periods t_0 and t_1 into the portions that are explained and unexplained by the counterfactual. Panels A and B show the decomposition for the true counterfactual estimated using the known data generating process. Panels C and D show the decomposition for the counterfactual estimated using DFL method. In the figure, we have defined a set of "natural" boundaries delineated by a 100×100 grid. Within each square in the grid, we use a uniform kernel to determine the sample density. This is equivalent to calculating the average density in the square, much like would be the case if we mapped average density by natural boundaries such as census block groups, tracts, or counties.

FIGURE 3: Counterfactual Spatial Distribution – Changes in Distribution Explained and Unexplained by Changes in X.

in parenthesis is the portion of the changes that can be explained by differences in the distribution of covariates.

Continuing with the parametric counterfactual example from Equation (6) shown in Figure 2, we decompose the changes in Y into those that can be explained by changes in X and those that cannot. As before, we have plotted the changes within a 100×100 grid of squares. In each square, the density is averaged over the area using a uniform kernel.

The results for the true counterfactual and the DFL estimate are shown in Figure 3. Panels A and C in Figure 3 show the portion of the changes in Y accounted for by changes in X in each case (from changes in σ). Panels B and D show the unexplained portion of the changes in Y that were due to the change in the parameter θ from the simulation. As we cannot reject that the true and DFL counterfactuals are the same distribution, it should not be surprising that the DFL counterfactual provides a very good spatial estimate of the true values of these changes.

The next step is to calculate a summary index to quantify how much of an observed change can be explained by the counterfactual over a given area, rather than at a single point. We choose to do so as follows. Let $G(\hat{f}_0, \hat{f}_1)$ be a statistic of interest that summarizes

© 2016 Wiley Periodicals, Inc.

the comparison of two densities f_0 and f_1 over an area A. We select a simple class of functions where:

$$G\ (f_0, f_1) = \iint\limits_A g(f_0(y_1, y_2), f_1(y_1, y_2))dy_1dy_2 \ = G_{01}, \tag{8}$$

and the explained share of G is:

$$Explained\ Share\ = \frac{G(f_{x^1 \rightarrow x^0}, f_1)}{G(f_0, f_1)} \ = \frac{G_{C1}}{G_{01}}. \tag{9}$$

G_{01} compares the observed period 0 and period 1 distributions. G_{C1} is the explained part of G_{01} because it compares the counterfactual distribution, which holds period 1 unobservables fixed with period 0 observables, to the observed period 1 distribution.

We have left this discussion general, as there are many potential statistics of interest. For example, we could be interested in the change in share explained at a particular geographic level, such as tract or county. This would be straightforward as $g\ (f_0(y_1, y_2), f_1(y_1, y_2)) = f_1\ (y_1, y_2) - f_0(y_1, y_2)$ at each point (y_1, y_2) in the area A that defines the tract or county. We could also be interested in other measures of change, such as aggregate change in share independent of direction (a measure of flux), where $g\ (f_0(y_1, y_2), f_1(y_1, y_2)) = |f_1(y_1, y_2) - f_0(y_1, y_2)|$. We could be interested in measures of concentration or dispersal (such as an index of suburbanization), where g is defined accordingly. This approach could also be extended to evaluate the explained share of other summary indices calculated from distributions, such as measures of segregation.

In order to quantify the share of the change explained, we create “tracts” and “counties” from our parametric example. We create the “tracts” by dividing the area into 625 squares in a 25×25 grid and the “counties” by dividing the area into 25 squares in a 5×5 grid. For each tract and county, we calculate the share explained by the counterfactual under the function $g\ (f_0(y_1, y_2), f_1(y_1, y_2)) = f_1\ (y_1, y_2) - f_0(y_1, y_2)$. In this case, we calculate the explained change in the density of individuals that live in a given tract or county over the total change in that location.

In Table 1, we report some basic summary statistics on the tract or county level distribution of shares explained. In the median tract, 20 percent of the observed change is accounted for by the counterfactual. In the median county, the share explained is 24 percent. On average, the share explained is 8 percent at the tract level and 29 percent at the county level.[8]

In addition to our baseline parametric example, we also tested two alternative parameterizations: (a) only the observables X change (determined by σ_{μ_j}, which changes from 0.5 to 1) and (b) only the unobservables ϵ_1^j and ϵ_2^j change (determined by θ_j, which changes from 0.2 to 0.5). We calculated the same explained share at the constructed tract and county levels.

With only the observables changing, at both the tract and county level, the median explained share is 101 percent. We also conducted the χ^2 tests comparing the distributions and could not reject that the counterfactual distributions (true or DFL estimate) were the same as the t_0 distribution. This is what we would expect given that the counterfactual holds the observables fixed at their t_0 distribution and the unobservables were unchanged in the model.

[8] Both the mean and median are calculated with each geography weighted by period 1 population. The means calculated are from the Windsorized distribution at the 1st and 99th percentile to limit the impact of extreme outliers.

© 2016 Wiley Periodicals, Inc.

TABLE 1: Percent of Change in Spatial Distribution in Parametric Example: Explained and Unexplained

A. "Tracts" – 25x25 grids		
Statistic	% Explained	% Unexplained
Mean	8.4	91.6
Distribution of Percent Explained		
10th Percentile	−100.7	200.7
25th Percentile	−59.4	159.4
50th Percentile	20.4	79.6
75th Percentile	34.4	65.6
90th Percentile	97.4	2.6
B. "Counties" – 5×5 grids		
Mean	29.3	70.7
Distribution of Percent Explained		
10th Percentile	−61.1	161.1
25th Percentile	−29.1	129.1
50th Percentile	23.5	76.5
75th Percentile	100.3	−0.3
90th Percentile	104.2	−4.2

Notes: This table shows descriptive statistics on the share of the change in the spatial distributions in the parametric example that can be accounted for by the change in observable and unobservable characteristics. For example, in the median "tract" 20 percent of the changes in the density can be accounted for by changes in X, while 80 percent remain "unexplained." We have grouped observations into "tracts" by dividing the area into 625 squares in a 25×25 grid and "counties" by dividing the area into 25 squares in a 5×5 grid. Both the mean and median are calculated with geographies weighted by the average density of individuals in period 1. The means shown are from the Windsorized distribution at the 1st and 99th percentile to limit the impact of extreme outliers.

With only the unobservables changing, the explained share at the tract level is 0 and 0.1 percent at the county level. In this case, the χ^2 tests did not reject that the counterfactuals differ from the period t_1 distribution. Again, this is as expected since the observables were unchanged from t_0, but the unobservables changed, and the counterfactual holds the unobservables fixed at their t_1 levels.

3. APPLICATION: SPATIAL DISTRIBUTION OF HISPANICS IN WASHINGTON, DC

To illustrate the application of the techniques developed in the previous section, we study the residential location choices of Hispanics in the Washington, DC,[9] metropolitan area using data from the U.S. Census and the National Historic Geographic Information System (NHGIS).[10] Understanding the location choices of households is at the core of urban and regional economics. Counterfactual spatial distributions can provide interesting *reduced-form* insights about the determinants of these choices. For the sake of brevity, in the analysis below we focus on *illustrating* the application of our methods and will remain somewhat agnostic about implications with the broader literature.

The residential location of minorities has been the focus of a large literature. While most papers in the literature study the determinants of urban segregation between blacks

[9] We focus on the Washington DC Core Based Statistical Area (CBSA). We will refer to the Washington DC CBSA as Washington DC or DC for brevity.

[10] The geographic data was provided by the NHGIS. It can be found on their website at http://www.nhgis.org. (Minnesota Population Center).

© 2016 Wiley Periodicals, Inc.

and whites (Reardon and O'Sullivan, 2004; Cutler, Glaeser, and Vigdor, 1999; Galster and Cutsinger, 2007, among many others), a growing number of studies analyze the patterns of segregation between Hispanics and other groups (Bayer, McMillan, and Rueben, 2004; Johnston, Poulsen, and Forrest, 2007; Iceland, 2004, for example).

Previous studies suggest that demographic characteristics are strongly associated with minority segregation. For example, Bayer et al. 2004 use 1990 census microdata in the San Francisco Bay area to show that sociodemographic characteristics account for much of the observed segregation of Hispanics and whites. Iceland and Wilkes 2006 explore the relationship between socioeconomic status and segregation and find that low status Hispanics are much more segregated from whites than high status Hispanics. They also found that while black-white segregation declined between 1990 and 2000, Hispanic-white segregation did not, in large part due to improvements in the socioeconomic status of blacks that did not occur for Hispanics. Chiswick and Miller 2005 discuss how ethnic goods and the lack of English language proficiency can cause immigrants to settle in concentrated or segregated areas of cities. Consistent with this model, Lichter et al. 2010 show that Hispanic-white segregation grew between 1990 and 2000 in the fastest growing Hispanic communities, greatly exceeding segregation levels in established ones. Iceland and Scopilliti 2008 found similar results when looking at segregation of recent immigrants. Hence, if the demographic characteristics of a minority such as the Hispanic population change over time, one would expect to observe changes in their spatial distribution.

In the rest of this section, we analyze by how much *changes* in the characteristics of Hispanics account for *changes* in the spatial distribution of Hispanics over time, and will focus our attention on the Washington DC area. Washington DC has a fast growing Hispanic population, both in absolute and relative terms. The Hispanic population grew from 217,000 in 1990 to 763,000 in 2010,[11] from 5.2 percent of the population to 13.9 percent. This growth in the Hispanic population has accompanied changes in the distribution of a variety of characteristics, including education relative to blacks and whites, age, citizenship, marriage rates, and household size.

Unconditional Spatial Distribution

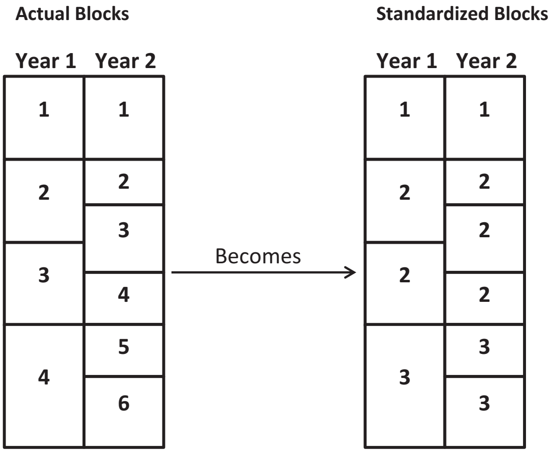
To compute the distribution of residential location choices, we use microdata from the 1990 decennial census long-form questionnaires and the 2012 five-year ACS file (with data from 2008 to 2012). These files identify the census block in which each surveyed household lives as well as individual income and demographic characteristics. The 1990 long-form questionnaires was completed by about one-sixth of all households in the United States and the ACS five-year file covers about one-eighth of all households.

To compute the spatial distribution of Hispanics across DC, we calculate the share of Hispanics/square mile in each census tract. However, over time the block and tract boundaries change. In order to have constant geographies, we use the block relationship files provided by the U.S. Census Bureau.[12] For blocks with changed boundaries, we group the smallest set of blocks in each period so that there are no overlaps in the boundaries between any set of standardized blocks across the two periods. An example of this process is shown in Figure 4. We then assign each set of standardized blocks into the 2010 census

[11] For ease of exposition, we refer to the 2012 five-year ACS file as 2010.

[12] Available at https://www.census.gov/geo/maps-data/data/relationship.html. The 2000 relationship files relate the blocks from 1990 to 2000 and the 2010 relationship files relate blocks from 2000 to 2010. We link these files to create the relationships from 1990 to 2010.

© 2016 Wiley Periodicals, Inc.



Notes: This figure shows a simple example of the block relationship mapping used to construct consistent geographies in the 2010 DC CBSA. It contains examples of the various types of blocks that are aggregated into larger groups of blocks that have the same boundaries in both periods.

FIGURE 4: Block Relationship Mapping Example.

tract in which the majority residents in the 2010 blocks live. In 98 percent of cases for both years, all blocks in the standardized group are in the same tract.[13]

We define the DC metro area as the current Core Based Statistical Area (CBSA). Our sample includes all individuals in a given year that reside in blocks whose centroid is within the boundaries of the DC CBSA. In the DC CBSA, there are 47,367 blocks in the 1990 census, 52,157 blocks in the 2000 census, and 91,418 blocks in the ACS. Using the standardization technique shown in Figure 4, matching the 1990–2010 blocks results in 29,740 standardized blocks.

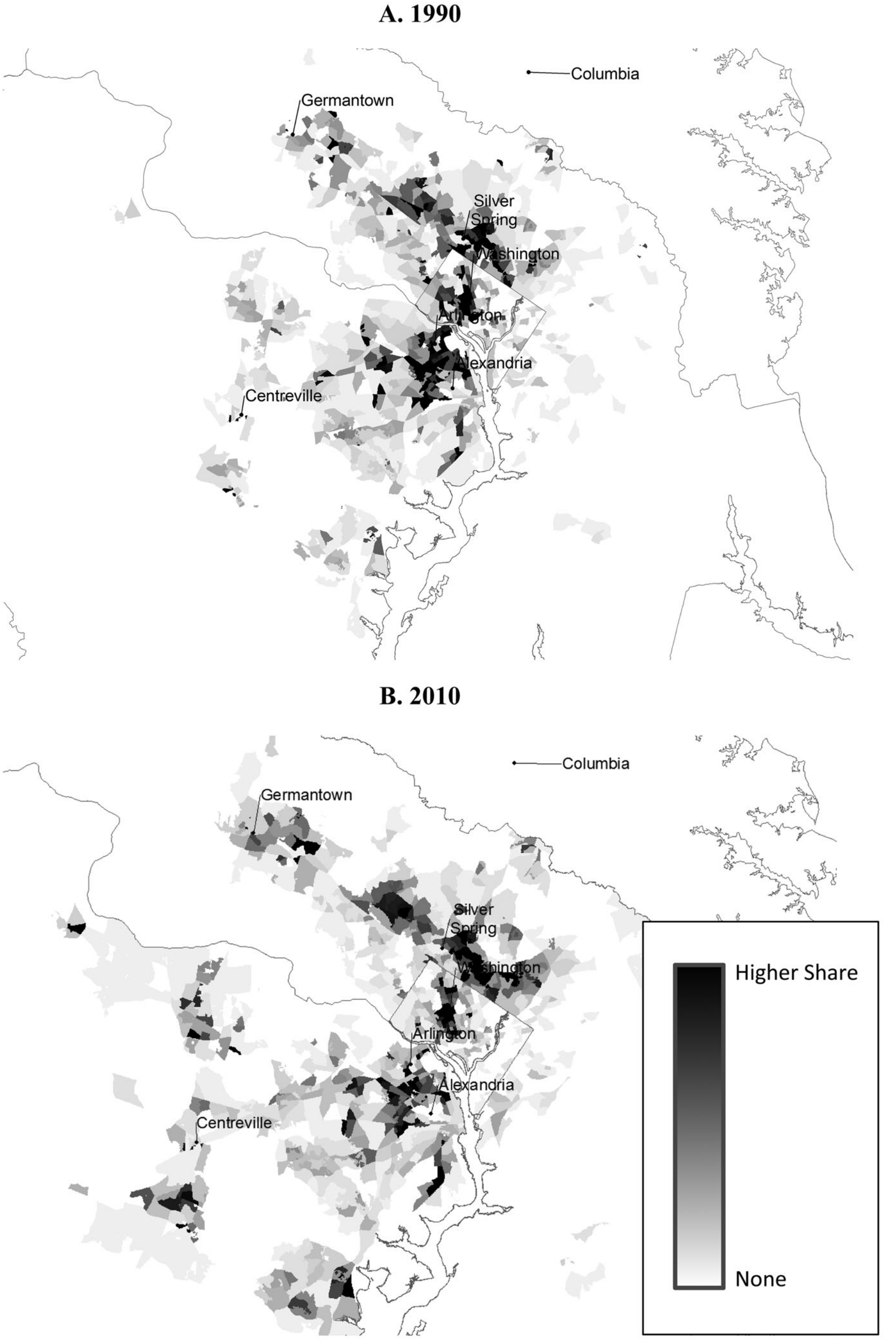
The DC CBSA includes 580,205 surveyed individuals in the 1990 census representing a population of 4.2 million people. In the 2012 five-year ACS, there are 371,156 observations representing 5.5 million people.

The survey data is used to compute the spatial distribution of Hispanics in DC for 1990 and 2010 by 2010 census tract, shown in Figure 5. In both periods, the highest share of Hispanics/square mile lives in northwest Washington DC and in the suburbs of Silver Spring, Arlington, and Alexandria. Figure 6 shows the change in the Hispanic distribution over this period. The share/square mile of Hispanics has decreased in many of the most dense areas and increased in other DC suburbs, including Prince George's County (southeast of Silver Spring), Germantown, etc.

During the same period, there were changes in the characteristics of Hispanics living in the Washington DC CBSA. Table 2 shows how the characteristics of Hispanics, along with blacks and whites, have changed over this period. With the growth in the Hispanic population, the age distribution of Hispanics has changed, with more children and individuals aged 56 and older. The share of citizens also changed, as a greater share of DC's Hispanics were born as citizens and fewer were not citizens in 2010. Mirroring changes for blacks and whites, fewer Hispanics 25 and over were married in 2010. Hispanics live in larger households in 2010 than in 1990. Hispanic household income also decreased relative to the CBSA average. Although not strictly a change, while Hispanic education

[13] The standardization procedure is discussed in greater detail in Appendix B.

© 2016 Wiley Periodicals, Inc.



A. 1990

B. 2010

FIGURE 5: Distribution of Hispanics in DC CBSA.

© 2016 Wiley Periodicals, Inc.

TABLE 2: Demographic Characteristics of DC CBSA population by race

	Whites (Non-Hispanic)			Blacks			Hispanics		
Variable	1990	2000	2010	1990	2000	2010	1990	2000	2010
Population (in Thousands)	2,719	2,820	2,700	1,033	1,275	1,416	217	417	763
(percent of CBSA)	(65.0)	(58.2)	(49.2)	(24.7)	(26.3)	(25.8)	(5.2)	(8.6)	(13.9)
Average Household Income*	115,005	132,504	153,249	73,951	82,591	92,552	82,987	88,797	93,518
(percent of CBSA Average)	(112.2)	(116.0)	(120.8)	(72.1)	(72.3)	(72.9)	(80.9)	(77.8)	(73.7)
Age (percent in each range)									
0-17	22.4	23.8	21.5	27.1	29.6	26.6	27.2	30.1	30.8
18–25	11.2	7.8	8.8	13.3	9.6	10.8	18.3	15.8	13.2
26–55	49.5	49.4	45.0	46.1	47.0	44.2	47.5	47.7	47.5
56–65	8.0	9.2	13.0	6.6	7.0	10.3	4.1	3.8	5.2
>65	8.8	9.8	11.6	6.8	6.8	8.1	2.9	2.6	3.2
Education** (percent in each group)									
< High School	10.4	7.0	3.8	25.4	17.4	9.9	34.9	41.7	35.8
High School	46.1	42.3	37.0	54.7	57.6	59.1	40.9	37.0	40.9
College	24.6	27.3	29.9	12.5	15.5	17.8	13.6	11.8	13.9
Graduate	18.8	23.4	29.2	7.5	9.5	13.1	10.5	9.6	9.4
Citizenship (percent in each group)									
Born a citizen	95.0	93.6	93.5	93.9	90.0	87.1	38.7	38.3	44.5
Naturalized	2.3	3.1	3.7	1.5	3.7	6.1	12.7	15.4	15.3
Not a citizen	2.7	3.3	2.8	4.6	6.3	6.8	48.7	46.3	40.2
Average share of noncitizens in HH	3.0	3.6	3.2	4.6	6.4	6.9	46.1	44.6	38.9
Marital Status** (percent in each group)									
Currently Married	67.0	66.3	63.7	44.5	45.1	41.7	62.4	62.6	56.2
Ever married	83.9	83.8	80.9	72.4	71.3	67.1	77.8	77.5	71.2
Never married	16.1	16.2	19.1	27.6	28.7	32.9	22.2	22.5	28.8
Married couple present in HH	72.8	71.6	69.4	48.7	47.6	45.5	70.1	71.5	65.3
Gender (percent in each group)									
Male	49.2	49.2	49.3	46.3	46.0	46.1	51.3	52.1	51.5
Female	50.8	50.8	50.7	53.7	54.0	53.9	48.7	47.9	48.5

(Continued)

© 2016 Wiley Periodicals, Inc.

TABLE 2: Continued

	Whites (Non-Hispanic)			Blacks			Hispanics		
Variable	1990	2000	2010	1990	2000	2010	1990	2000	2010
Household Size (percent in each group)									
1 Person	9.8	11.2	12.1	10.2	10.9	12.4	4.3	3.5	3.8
2 People	26.5	27.9	29.0	19.0	20.4	22.3	13.8	10.4	11.5
3 People	21.3	19.4	19.5	21.6	21.6	20.9	17.4	14.9	16.7
4 People	23.9	23.1	22.3	21.2	20.9	20.6	22.9	21.7	24.6
5+ People	18.4	18.4	17.2	28.0	26.2	23.8	41.6	49.6	43.3
Household with: (percent in each group)									
Children	50.3	51.6	47.3	60.4	62.5	56.7	67.6	73.3	71.7
Individuals 65+	13.9	15.0	18.2	14.5	14.4	17.3	8.3	8.4	9.3

Source: U.S. Census internal microdata from 1990 and 2000 longform census and 2012 five-year ACS file. The ACS five-year file data is from 2008–2012 and is labeled 2010 in the table.

Notes: *Household income in 2012 dollars. Total household income is assigned to each individual and then the average is taken across all individuals in the CBSA. **Educational attainment and marital status for individuals over 25.

© 2016 Wiley Periodicals, Inc.

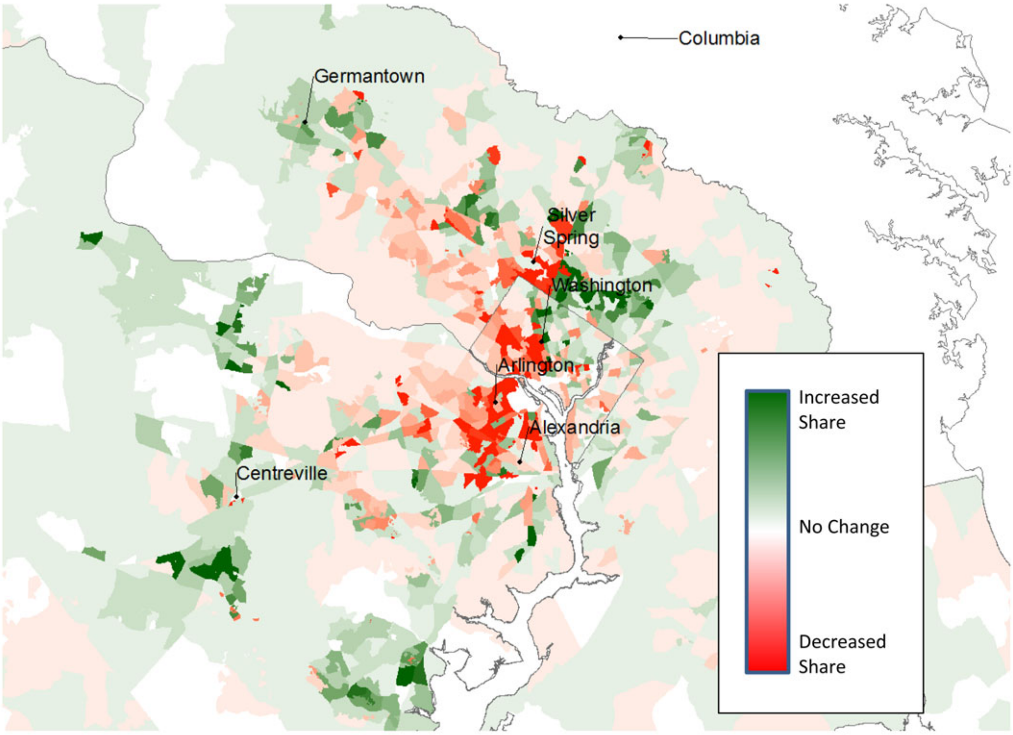


FIGURE 6: Change in Distribution of Hispanics in DC CBSA.

levels are relatively unchanged from 1990 to 2010 (with a decline in 2000), for blacks and whites, education levels increased steadily over time, with fewer non-high school graduates and more college and graduate educated individuals.

How do changes in these underlying characteristics of age, citizenship, marital status, and income account for changes in the spatial distribution of Hispanics? To answer this question, we apply the methods developed in the previous section to estimate how the distribution of Hispanics in 2010 would look if their demographic characteristics had remained constant at their 1990 distributions.

Counterfactual Spatial Distribution

To calculate the counterfactual distribution, we estimate the logistic regression on the variables shown in Table 3 on the pooled sample of 1990 census and 2012 five-year ACS observations. As explained in Section 2, the dependent variable is a dummy indicating whether the observation was surveyed in the 1990 census (value of 1) or the ACS (value of 0). We report the coefficients from the regression in the table. Positive coefficients indicate these characteristics were more prevalent amongst Hispanics in DC in 1990. As expected from the summary stats in Table 2, income, family size, marital status, age, and lack of citizenship are predictive of an observation being in the 1990 census sample.

We calculate a pseudo-R^2 value for the logistic regression, the Tjur-R^2 (Tjur, 2009). The Tjur-R^2 is simple to calculate and easy to interpret. It is computed by averaging the logistic regression predictions separately for each of the two categories of the dependent variable (1 and 0) and taking the difference between the means. If the regression model

© 2016 Wiley Periodicals, Inc.

TABLE 3: Logistic Regression Results: Reweighting 2010 Hispanics to 1990 Characteristics

Variable	Coefficient (SE)	Variable	Coefficient (SE)
Income Relative to Mean (Log)	0.89	Age	
	(0.04)	0-17	0.74
Family Size			(0.05)
1 Person	0.37	18–25	0.81
	(0.06)		(0.07)
3 People	−0.27	26–35	1.10
	(0.04)		(0.10)
4 People	−0.49	36–45	1.38
	(0.04)		(0.13)
5 People	−0.44	46–55	2.04
	(0.05)		(0.16)
6 People	−0.52	56–65	2.77
	(0.05)		(0.19)
7 People	−0.55	66–75	3.28
	(0.06)		(0.23)
8 People	−1.21	Age (Continuous)	−0.07
	(0.07)		(0.00)
9 People	−1.27	Children in HH	0.01
	(0.09)		(0.03)
10 People	−1.42	Aged 65+ in HH	0.07
	(0.10)		(0.04)
Education		Citizenship	
1-8 Grade or less	0.05	Born in U.S. Territory (Not in U.S. State)	1.12
	(0.03)		(0.06)
9–12 (No HS Diploma)	−0.18	Born Abroad to U.S. Parents	0.75
	(0.04)		(0.07)
HS Diploma	0.06		
	(0.04)	Naturalized	0.29
Some College	−0.02		(0.04)
	(0.05)	No	0.33
Bachelor's Degree	−0.05		(0.04)
	(0.06)	Share noncitizens in HH	0.65
			(0.04)
Marital Status		Married couple in HH	0.20
Married	0.63		(0.03)
	(0.04)	Male	−0.03
Widowed	1.35		(0.02)
	(0.09)	Max Education in HH	
Divorced	0.50	1-8 Grade or less	−0.03
	(0.06)		(0.06)
Separated	0.77	9-12 (No HS Diploma)	0.20
	(0.07)		(0.05)
Constant	−1.01	HS Diploma	0.02
	(0.06)		(0.04)
		Some College	0.07
			(0.04)
		Bachelor's Degree	−0.10
			(0.04)

(Continued)

© 2016 Wiley Periodicals, Inc.

TABLE 3: Continued

Variable	Coefficient (SE)	Variable	Coefficient (SE)
Observations	65,476		
Tjur $\boldsymbol{R}^2$	0.06		

Notes: This table reports the logit coefficients for the DFL reweighting regression. The sample includes all Hispanics in the DC CBSA surveyed in the 1990 census longform (42 percent of the sample) or the 2012 five-year ACS file (58 percent of the sample). The dependent variable is equal to 1 for the 1990 sample and 0 for the ACS sample. Therefore a positive coefficient (such as the 0.33 on not a citizen) means that conditional on the other characteristics, noncitizens are more likely to be in the 1990 census sample than the ACS sample.

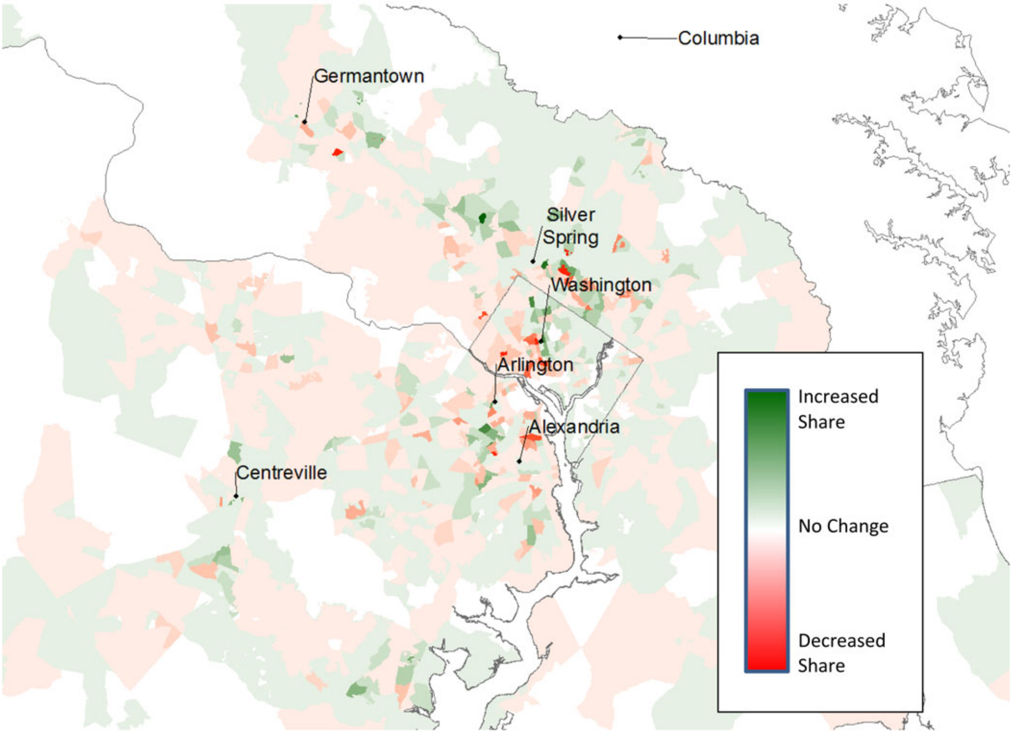


FIGURE 7: Explained Portion of Change in Distribution of Hispanics in DC CBSA.

predicts the dependent variable well, the Tjur-R^2 increases. The Tjur-R^2 is bounded between 1 (perfect prediction of the dependent variable) and 0 (the model offers no value in predicting the dependent variable). In this case, the Tjur-R^2 is relatively low at 0.06.[14]

We then decompose the change in the distribution of Hispanics using Equation (7). We map the explained and unexplained portions of the change in Figures 7 and 8, respectively, both shown in the same scale as Figure 6. While the counterfactual seems to account for the general direction of the changes in the distribution of Hispanics, it appears that nearly

[14] It should be noted that a very high R^2 would also be problematic for a DFL counterfactual as the technique requires that X values in the support of the period 0 distribution also be in the support of the period 1 distribution.

© 2016 Wiley Periodicals, Inc.

TABLE 4: Percent of Change in Hispanic Distribution Explained and Unexplained

A. 2010 Tract

Statistic	% Explained	% Unexplained
Mean	15.4	84.6
Distribution of Percent Explained		
10th Percentile	−23.5	123.5
25th Percentile	−8.5	108.5
50th Percentile	11.0	89.0
75th Percentile	22.9	77.1
90th Percentile	52.8	47.2

B. Counties

		Percent of Hispanic Population			
County	State	% Explained	% Unexplained	1990	2010
Charles County	MD	−86.5	186.5	0.7	0.8
Jefferson County	WV	−15.8	115.8	0.2	0.3
Montgomery County	MD	−11.6	111.6	24.4	21.6
Falls Church City	VA	−11.6	111.6	0.2	0.1
Fairfax City	VA	−9.6	109.6	0.5	0.4
Loudoun County	VA	−8.7	108.7	0.8	4.6
Frederick County	MD	0.2	99.8	0.7	2.5
Arlington County	VA	2.5	97.5	10.3	4.0
Prince George's County	MD	3.7	96.3	12.3	16.0
Fredericksburg City	VA	4.0	96.0	0.2	0.3
Prince William County	VA	4.7	95.3	4.1	11.6
Spotsylvania County	VA	5.4	94.6	0.4	1.2
District of Columbia	DC	5.4	94.6	13.6	7.1
Stafford County	VA	9.5	90.5	0.5	1.5
Alexandria City	VA	13.5	86.5	4.6	2.7
Manassas City	VA	14.9	85.1	0.4	0.9
Fauquier County	VA	15.2	84.8	0.3	0.6
Manassas Park City	VA	17.4	82.6	0.1	0.5
Clarke County	VA	29.9	70.1	0.0	0.1
Warren County	VA	66.0	34.0	0.1	0.2
Fairfax County	VA	66.1	33.9	23.0	22.3
Calvert County	MD	80.1	19.9	0.2	0.3
Average		14.1	85.9		
10th Percentile		−11.6	111.6		
25th Percentile		−8.7	108.7		
Median		3.7	96.3		
75th Percentile		13.5	86.5		
90th Percentile		66.1	33.9		

Notes: This table shows basic statistics on the percent of change in the Hispanic distribution explained by the change in demographic characteristics in the regression in Table 3. The tracts and counties are weighted by share of Hispanics/square mile in 2010 for both the mean and distribution of percent explained. For the median tract, 11 percent of the change in the distribution is explained by the change in demographic characteristics and 89 percent is not. For the median county, 4 percent is explained and 96 percent is not. The weighted average explained is 15 percent at the tract level and 14.1 percent at the county level. There are 378 tracts and 22 counties with Hispanics in the DC CBSA.

© 2016 Wiley Periodicals, Inc.

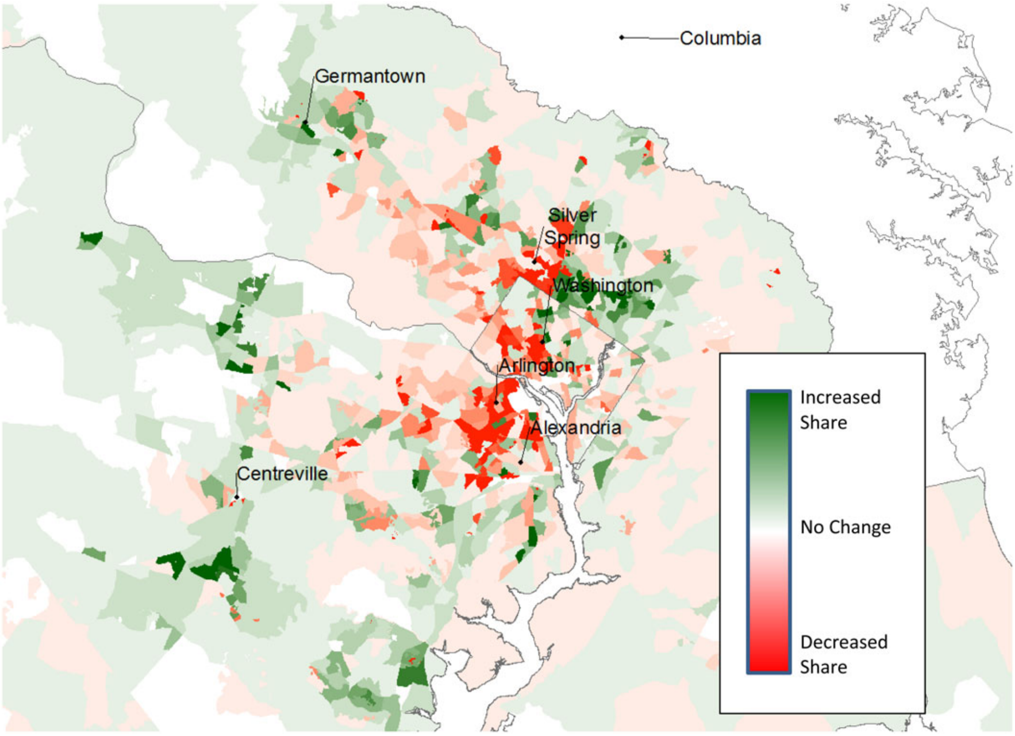


FIGURE 8: Unexplained Portion of Change in Distribution of Hispanics in DC CBSA.

all of the change is unexplained.[15] In Figure 7, the change in observable characteristics of Hispanics predicts a relatively small change in the distribution of Hispanics at the tract level, but from Figure 6 we see much larger changes in the distribution.[16]

As in the parametric example, we quantify the explained change for each tract and county, with some summary results shown in Table 4. We also calculate the average of these percentages, weighted by the 2010 Hispanic population in the geographic unit. At the tract level, an average of 15 percent of the change can be accounted for by the change in the demographic characteristics over the period and 85 percent is unaccounted for by the counterfactual. At the county level, the weighted average of the explained share is 14 percent. We also provide some distributional statistics on the explained and

[15] Note that in our model, individuals with characteristics X are free to choose their location in each period, and their current location choices are assumed to be independent of their past location. For instance, if young individuals in period t=0 become old in period t=1, our setup assumes that they are free to choose where to live in period 1 without regard to their location in period 0. This is an important limitation of our current set up, and should be taken into consideration when interpreting our results. We are not sure how to introduce dependence in the time-space dimension within the context of the DFL decomposition. This is certainly an important topic that deserves future research.

[16] We compared the 1990, 2010, and counterfactual distributions to each other using the χ^2 goodness of fit test used in Section 2. In all cases, we can reject the distributions are the same with a P-value of virtually zero.

© 2016 Wiley Periodicals, Inc.

unexplained change. At the median tract, 11 percent of the change in the spatial distribution is explained. At the median county, only 4 percent is explained.[17]

Access to Census individual microdata is generally limited. For example, the public-use microdata from the decennial censuses and the ACS do not include geographic identifiers at the tract or even the county level. Instead, block, block group, and tract level data are available aggregated by group with varying levels of information. The methods developed in this paper can also be applied in such cases when only aggregate data are available. For the sake of brevity, we leave the discussion of such application to Appendix C, and highlight here that decomposition results are similar if individual level or aggregate data are used.

4. CONCLUSION

The paper is a first attempt to extend the notion of counterfactual distributions to spatial problems. We apply the DFL decomposition technique to construct counterfactual spatial distributions. Using a simple simulation, we show that the spatial counterfactual technique can provide an accurate decomposition of the extent to which changes in location decisions are due to changes in a particular underlying set of characteristics. We then illustrate how counterfactual spatial distributions can be used to understand how individual characteristics have affected the location decisions of Hispanics in the Washington DC area using survey data from the 1990 census and 2008–2012 ACS.

We hope our approach is useful in other applications.

APPENDIX A: PARAMETRIC EXAMPLE MODEL

In this section, we provide details about the parametric example in Section 2. The location coordinates chosen by individual i in period j are given by $y^j_{1i} = \epsilon^j_{1i}\, X^j_i$ and $y^j_{2i} = \epsilon^j_{2i}\, X^j_i$, where time period $j = \{0, 1\}$. We assume that X is a random variable and that its distribution depends on a sole parameter $\sigma_{\mu j}$. In particular, we let $X^j = \frac{\exp\{\mu^j\}}{1+\exp\{\mu^j\}}$, and $\mu^j \sim N(0, \sigma_{\mu j})$.

We assume that ϵ^j_1 and ϵ^j_2 are random variables independent of X. For instance, we let $\epsilon^j_k = \pi^j\, W_k + (1 - \pi^j)Z_k$, $W_k \sim N(\mu_{Wk}, \sigma_{Wk})$, $Z_k \sim N(\mu_{Zk}, \sigma_{Zk})$, $k = \{1, 2\}$, and $\pi^j \sim$ *Bernoulli*(θ_j). In this simple model, the only parameters that are time dependent are $\sigma_{\mu j}$ and θ_j. Notice that $\sigma_{\mu j}$ measures the dispersion of the covariates in each time period, while θ_j denotes heterogeneity in location preferences that are independent of X. Given a set of parameter values $\Omega = \{\mu_{W1}, \sigma_{W1}, \mu_{Z1}, \sigma_{Z1}, \mu_{W2}, \sigma_{W2}, \mu_{Z2}, \sigma_{Z2}, \sigma_{\mu j}, \theta_j\}$ one can simulate (say, 1,000,000) realizations of these random variables in each period j and estimate the joint density of $[y_1, y_2]$.

The parameter values that change over time are set to $\sigma_{\mu_0} = 0.5$, $\sigma_{\mu_1} = 1$, $\theta_0 = 0.2$, and $\theta_1 = 0.5$. The time invariant coefficients are set to $\mu_{W1} = -2$, $\sigma_{W1} = 1$, $\mu_{Z1} = 2$, $\sigma_{Z1} = 1$, $\mu_{W2} = 2$, $\sigma_{W2} = 1$, $\mu_{Z2} = -2$, and $\sigma_{Z2} = 1$.

[17] Unaccounted changes may be explained by changes in preferences, regulation environment, and other factors. For instance, Rothwell (2011) finds that restrictive or exclusionary local land regulation is an important component of observed segregation of blacks and Hispanics.

© 2016 Wiley Periodicals, Inc.

APPENDIX B: DETAILED CREATION OF 1990 CENSUS AND 2012 FIVE-YEAR ACS COUNTERFACTUAL

In this section, we discuss in greater detail how our samples and geographic data were put together and used to generate the spatial counterfactuals. The mapping and boundary shapefiles were provided by the National Historic Geographic Information System (NHGIS). With the census block files for all states in the DC CBSA (DC, Maryland, Virginia, and West Virginia) for 1990, 2000, and 2010, we selected only those blocks in each state and year with their centroid within the current CBSA boundaries. In the DC CBSA, there are 47,367 blocks in the 1990 census, 52,157 blocks in the 2000 census, and 91,418 blocks in the ACS. Using the block relationship files from the U.S. Census Bureau,[18] we constructed standardized blocks (S-blocks) relating: 1) 1990 to 2000 and 2) 2000 to 2010 separately. To create the S-blocks, for a given block in period 1, we did the following:

(1) Find all the blocks in period 2 that matched the initial period 1 block.

(2) Find any additional period 1 blocks that matched to the set of period 2 blocks from step 1.

(3) For the additional period 1 blocks repeat steps 1 and 2 until no new period 1 and 2 blocks are added to the set of matched blocks.

(4) Assign all matched period 1 and 2 blocks to a set of grouped blocks to create a single S-block.

(5) Repeat this for all period 1 blocks until all blocks in both periods are assigned to an S-block.

We then linked the 1990–2000 and 2000–2010 S-blocks as above to create ones that spanned the 1990 to 2010 geographies. Therefore, any block boundaries that changed between 1990–2000 and 2000–2010 were aggregated into the smallest set of grouped blocks that had the same boundaries in 1990 and 2010, for a total of 29,740 S-blocks.

In order to assign each S-block into a 2010 census tract, we calculated the share of the 2012 five-year ACS population in each S-block in each census tract and assigned the S-block in both periods to the tract with the highest share. For both periods, over 98 percent of S-blocks were in the same 2010 tract.

Finally, we created the extracts of individual observations from the 1990 census and 2012 five-year ACS file. This includes all individuals that resided in blocks within the current DC CBSA. We standardized all variables used to the 2012 five-year ACS categories and definitions and calculated the household characteristics for each individual. We then ran the regression shown in Table 3 using the person weights from each file. We created the counterfactual by multiplying the person weights in the 2012 five-year ACS file by the reweighting factor $\hat{\tau}_{t_{2010} \to t_{1990}}$ as calculated in the logistic regression. Finally, to generate the maps, we summed the relevant weights (person weights or counterfactual-adjusted) in each tract for Hispanics in each survey.

[18] Available at https://www.census.gov/geo/maps-data/data/relationship.html. The 2000 relationship files relate the blocks from 1990 to 2000 and the 2010 relationship files relate blocks from 2000 to 2010. We link these files to create the relationships from 1990 to 2010.

© 2016 Wiley Periodicals, Inc.

APPENDIX C: COUNTERFACTUAL SPATIAL DISTRIBUTIONS WITH AGGREGATED DATA

In many cases, the individual microdata is not publicly available. For example, the public-use microdata from the decennial censuses and the ACS do not include geographic identifiers at the tract or even the county level. Instead, block, block group, and tract level data is available aggregated by group with varying levels of information: block data is available broken down only by age groups and ethnicity; block group data contains information on education and average income.

In this section, we show an example of our counterfactual method using block group data from the 1990 census and 2008–2012 five-year ACS file. To do so, we will make the additional assumption that all Hispanics living inside an area have the same characteristics as those of a "representative individual" with the mean income and average or median other characteristics of the residents that live there. For example, if Hispanics in a block group earn a per capita income of $20,000 and the median age is 45, with a given education distribution, the representative Hispanic individual for that block group will have each of these characteristics. Given these assumptions, one can easily apply the methods developed in the previous section and estimate the counterfactual distributions.

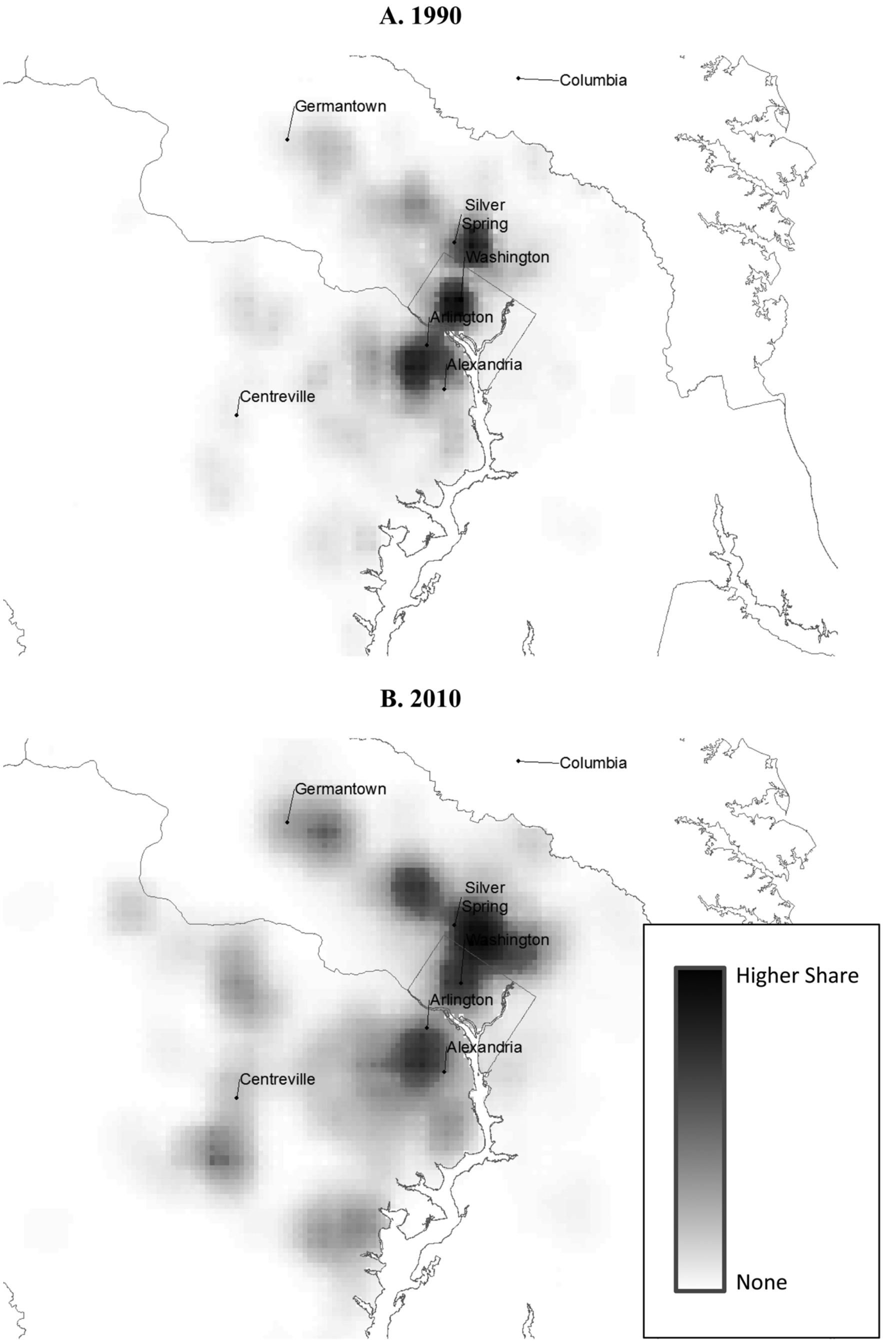
In this simple example, we include only median age (category), per capita income relative in the block group relative to the CBSA average, and share of education (< high school, high school, some college, and college or graduate degree).

In this example, we map the spatial distribution using a kernel density. We do so as it is another possible way, in contrast to aggregating individuals by tract, that the spatial data can be mapped. The distributions were estimated using a bivariate kernel with a 100×100 grid of points covering the DC CBSA.[19] The location of each representative individual is at the centroid of the block group in which they reside. Each individual is given a weight corresponding to the number of Hispanics in the block group.

We show the kernel density of the distribution of Hispanics in 1990 and 2010 using block group data, in Figures A1, which corresponds to Figure 5 at the standardized tract level. In Figure A2, we show the change in the distribution of Hispanics between 1990 and 2010, which corresponds to Figure 6. In both cases, the maps are very similar, whether drawn over the natural boundaries or using kernel density estimation from the block group centroids. In Figures A3 and A4, we show the change explained and unexplained by the counterfactual distribution. At the median point (weighted by the 2010 Hispanic density) in the kernel, 3.6 percent of the change is explained. In other words, with the smaller set of observables in this example, none of the change in the distribution of Hispanics can be explained by the changes in observables between 1990 and 2010. It is not surprising that we can explain a smaller share of the change with the block level aggregate data than with the individual microdata as we have fewer variables in the model in the aggregate case.

[19] The kernel was implemented using the bidensity stata program written by John Luke Gallup and Christopher Baum.

© 2016 Wiley Periodicals, Inc.



A. 1990

B. 2010

FIGURE A1: Distribution of Hispanics in DC CBSA.

© 2016 Wiley Periodicals, Inc.

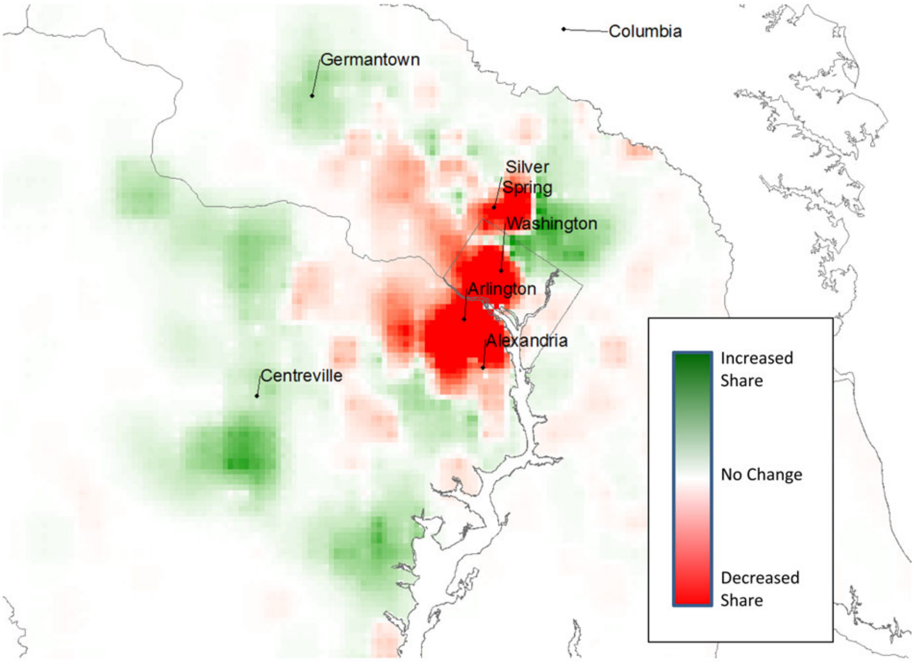


FIGURE A2: Change in Distribution of Hispanics in DC CBSA.

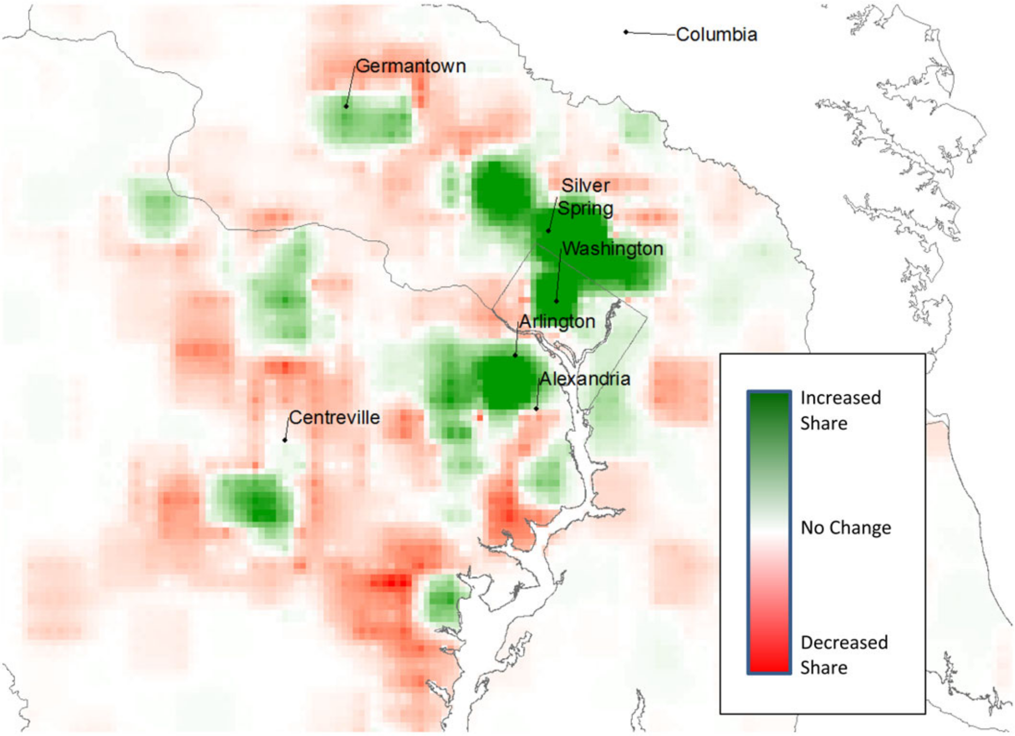


FIGURE A3: Explained Portion of Change in Distribution of Hispanics in DC CBSA.

© 2016 Wiley Periodicals, Inc.

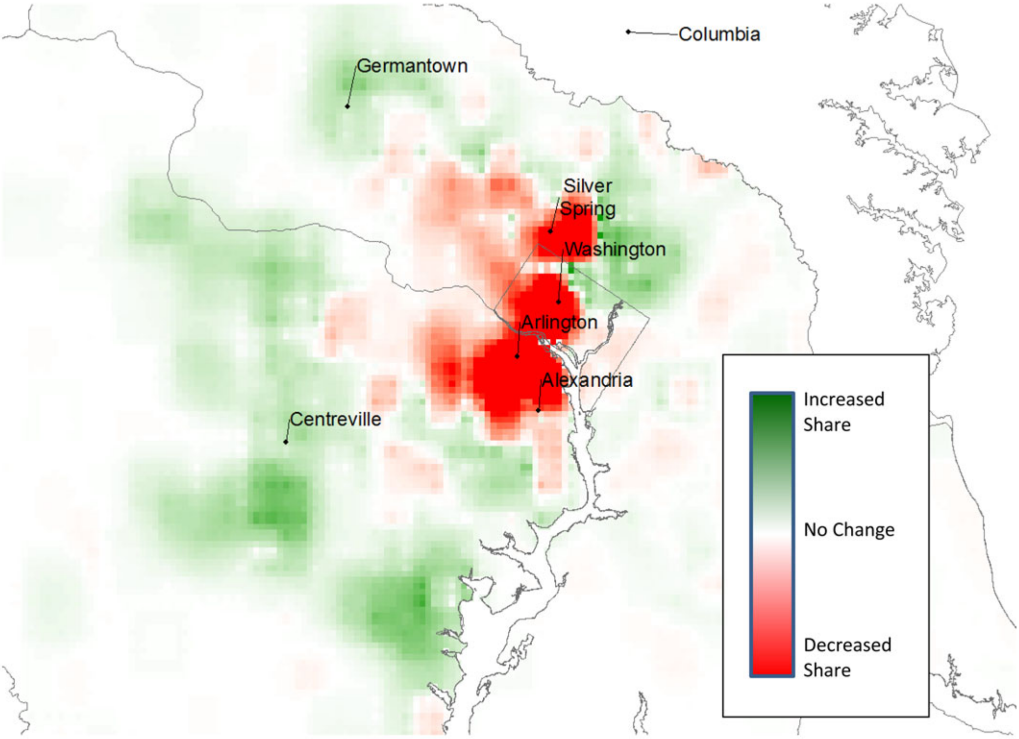


FIGURE A4: Unexplained Portion of Change in Distribution of Hispanics in DC CBSA.

REFERENCES

Albrecht, James, Anders Bjorklund, and Susan Vroman. 2003. "Is There a Glass Ceiling in Sweden?" *Journal of Labor Economics*, 21(1), 145–177.

Aslan, B. and Gunter Zech. 2005. “Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding, ” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 537(3), 626–636.

Bayer, Patrick, Robert McMillan, and Kim S. Rueben. 2004. “What Drives Racial Segregation? New Evidence Using Census Microdata,” *Journal of Urban Economics* 56(3), 514–535.

Bound, John, Michael F. Lovenheim, and Sarah Turner. 2009. “Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources,” *American Economic Journal: Applied Economics*, 2(3), 129–157.

Carrillo, Paul E. and Anthony Yezer. 2009. “Alternative Measures of Homeownership Gaps Across Segregated Neighborhoods,” *Regional Science and Urban Economics*, 39(5), 542–552.

Carrillo, Paul E. and Jaren C. Pope. 2012. “Are Homes Hot or Cold Potatoes? The Distribution of Marketing Time in the Housing Market,” *Regional Science and Urban Economics*, 42(1-2), 189–197.

Chiswick, Barry R and Paul W Miller. 2005. “Do Enclaves Matter in Immigrant Adjustment?” *City & Community*, 4(1), 5–35.

Cobb-Clark, Deborah A. and Mathias G. Sinning. 2011. “Neighborhood Diversity and the Appreciation of Native- and Immigrant-Owned Homes,” *Regional Science and Urban Economics*, 41(3), 214–226.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. “The Rise and Decline of the American Ghetto,” *Journal of Political Economy*, 107(3), 455–506.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. “Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach,” *Econometrica*, 64(5), 1001–1044.

Dolls, Mathias, Clemens Fuest, and Andreas Peichl. 2012. “Automatic Stabilizers and Economic Crisis: US vs. Europe,” *Journal of Public Economics*, 96(3), 279–294.

© 2016 Wiley Periodicals, Inc.

Donald, Stephen G., David A. Green, and Harry J. Paarsch. 2000. “Differences in Wage Distributions Between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates,” *Review of Economic Studies*, 67 (4), 609–633.

Duranton, Gilles and Henry Overman. 2005. “Testing for Localisation Using Micro-Geographic Data, ” *Review of Economic Studies*, 72(4), 1077–1106

——. 2008. “Exploring Detailed Patterns of UK Manufacturing Location Using Microgeographic Data,” *Journal of Regional Science*, 48(1), 213–243

Dustmann, Christian, Johannes Ludsteck, and Uta Schonberg. 2009. “Revisiting the German Wage Structure,” *The Quarterly Journal of Economics*, 124(2), 843–881.

Fasano, Giovanni and Alberto Franceschini. 1987. “A multidimensional version of the Kolmogorov–Smirnov test,” *Monthly Notices of the Royal Astronomical Society*, 225(1), 155–170.

Fesselmeyer, Eric, Kien T. Le, and Kiat Ying Seah. 2012. “A Household-Level Decomposition of the White–Black Homeownership Gap,” *Regional Science and Urban Economics*, 42(1-2), 52–62.

——. 2013. “Changes in the White–Black House Value Distribution Gap from 1997 to 2005,” *Regional Science and Urban Economics*, 43(1), 132–141.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. “Unconditional Quantile Regressions,” *Econometrica*, 77(3), 953–973.

Fortin, Nicole M., Thomas Lemieux, and Sergio Firpo. 2011. “Decomposition Methods in Economics,” *Handbook of Labor Economics*, 4(11), 1–102.

Galster, George and Jackie Cutsinger. 2007. “Racial Settlement and Metropolitan Land-Use Patterns: Does Sprawl Abet Black-White Segregation?” *Urban Geography*, 28(6), 516–553.

Iceland, John. 2004. “Beyond Black and White Metropolitan Residential Segregation in Multi-Ethnic America,” *Social Science Research*, 33(2), 248–271.

Iceland, John and Melissa Scopilliti. 2008. “Immigrant Residential Segregation in U.S. Metropolitan Areas, 1990-2000,” *Demography*, 45(1), 79–94.

Iceland, John and Rima Wilkes. 2006. “Does Socioeconomic Status Matter? Race, Class, and Residential Segregation,” *Social Problems*, 53(2), 248–273.

Johnston, Ron, Michael Poulsen, and James Forrest. 2007. “Ethnic and Racial Segregation in U.S. Metropolitan Areas, 1980-2000: The Dimensions of Segregation Revisited,” *Urban Affairs Review*, 42(4), 479–504.

Lemieux, Thomas. 2006. “Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?” *The American Economic Review*, 96(3), 461–498.

Lemieux, Thomas, W. Bentley MacLeod, and Daniel Parent. 2009. “Performance Pay and Wage Inequality,” *The Quarterly Journal of Economics*, 124(1), 1–49.

Lichter, Daniel T., Domenico Parisi, Michael C. Taquino, and Steven Michael Grice. 2010. “Residential Segregation in New Hispanic Destinations: Cities, Suburbs, and Rural Communities Compared,” *Social Science Research*, 39(2), 215–230.

Machado, José A. F. and José Mata. 2005. “Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression,” *Journal of Applied Econometrics*, 20(4), 445–465.

McMillen, Daniel P. 2008. “Changes in the Distribution of House Prices over Time: Structural Characteristics, Neighborhood, or Coefficients?” *Journal of Urban Economics*, 64(3), 573–589.

Minnesota Population Center. 2011. *National Historical Geographic Information System: Version 2.0*. Minneapolis, MN: University of Minnesota.

Nicodemo, Catia and Josep Maria Raya. 2012. “Change in the Distribution of House Prices Across Spanish Cities,” *Regional Science and Urban Economics*, 42(4), 739–748.

O’Donnell, Owen, Ángel López Nicolás, and Eddy Van Doorslaer. 2009. “Growing Richer and Taller: Explaining Change in the Distribution of Child Nutritional Status During Vietnam’s Economic Boom,” *Journal of Development Economics*, 88(1), 45–58.

Pallais, Amanda. 2009. “Taking a Chance on College Is the Tennessee Education Lottery Scholarship Program a Winner?” *Journal of Human Resources*, 44(1), 199–222.

Reardon, Sean F. and David O’Sullivan. 2004. “Measures of Spatial Segregation,” *Sociological Methodology*, 34(2004), 121–162.

Rothwell, Jonathan T. 2011. “Racial Enclaves and Density Zoning: The Institutionalized Segregation of Racial Minorities in the United States,” *American Law and Economics Review*, 13(1), 290–358.

Tjur, Tue (2009). “Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination,” *The American Statistician*, 63(4), 366–372

© 2016 Wiley Periodicals, Inc.