Counterfactual Spatial Distributions* 

Paul E. Carrillo a
George Washington University

Jonathan Rothbaum b
George Washington University

March 2014

Abstract

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 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 1) how much changes in individual characteristics of Hispanics in the Washington, DC, area account for changes in the distribution of their residential location choices, and 2) how changes in the average characteristics of shareholders account for changes in the spatial distribution of new firms in Quito, Ecuador.

Keywords: Decomposition, Non-parametric Estimation

JEL Codes: C14, R23, R30

* 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 and Urban Economics Association for helpful comments and discussions. We are grateful to the staff of the Centro de Estudios Fiscales of the Ecuadorian Tax Authority and the Ecuadorian National Institute of Statistics (INEC) for their outstanding collaboration and to the George Washington Institute for International Economic Policy (IIEP) for research support.

a Department of Economics and ESAI - George Washington University; email: pcarrill@gwu.edu.
b Institute for International Economic Policy - George Washington University; email: jrothbau@gwmail.gwu.edu.
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.\textsuperscript{1} These techniques have been used recently in several studies in urban, labor, public and development economics.\textsuperscript{2} 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. In the classical Oaxaca-Blinder wage decomposition researchers often simulate the counterfactual mean: how the mean wage of a demographic group would look like if they experience 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

\begin{footnotesize}
\begin{enumerate}
\item For a comprehensive review about decomposition methods in Economics see Fortin, Lemieux, & Firpo (2011).
\item For applications in labor, see Albrecht et al. (2003), Albrecht et al. (2009), Dustmann, Ludsteck, and Schonberg (2009), Lemieux, MacLeod, and Parent (2009), 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 and 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), Bargain et al. (2009), O’Donnell, López Nicolás, and Van Doorslaer (2009), and Nguyen et al. (2007).
\end{enumerate}
\end{footnotesize}
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 demographic characteristics (such as age, race and gender, 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, Fortin, and Lemieux (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 p.d.f. 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.

We then illustrate the spatial counterfactual technique to understand 1) how changes in individual characteristics of Hispanics have affected their location choices in the Washington, DC area and 2) how changes in the characteristics of shareholders affected the location choices of new firms in Quito, Ecuador. Our applications show that counterfactual spatial distributions are straightforward to compute and can provide interesting reduced form insights about the determinants of location choices of individuals and firms. The rest of the paper is structured as follows. The next section presents the econometric methodology. In the third section we present our empirical applications. The last section concludes.

2. Methods

2.1 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 in 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, income, and race, for example).

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

\[
1 \quad f(y_1, y_2|T = t_1) = \int f(y_1, y_2|x, T = t_1)h(x|T = t_1)dx,
\]

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 population in period \( t_0 \) is defined similarly.

\[
2 \quad f(y_1, y_2|T = t_0) = \int f(y_1, y_2|x, T = t_0)h(x|T = t_0)dx.
\]
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 non-parametric (kernel) method.

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

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

### 2.2 Parametric example

The construction of the counterfactual spatial distribution in equation (3) can be straightforward if the data generating process were known. To illustrate this, we specify a simple parametric model (described in more detail in the Appendix) of the location coordinates $y_1$ and $y_2$ based on a single characteristic $X$ and "taste" parameters $\epsilon_2^j$ and $\epsilon_2^j$. Formally, the location coordinates chosen by individual $i$ in period $j$ are given by $y_{1i}^j = \epsilon_{1i}^jX_i^j$ and $y_{2i}^j = \epsilon_{2i}^jX_i^j$, where time period $j = \{0,1\}$. We assume that $X$ is a random variable and that its distribution depends on a sole parameter $\sigma_{\mu_j}$ which measures the dispersion of the $X$ covariate in each time period. We also let parameters $\epsilon_1^j$ and $\epsilon_2^j$ be random and their distribution depend only on parameter $\theta_j$.

Given a set of parameter values and the assumption that $\sigma_{\mu_1} > \sigma_{\mu_2}$ and $\theta_1 > \theta_2$, we simulate 10,000 realizations of these random variables in each period and estimate the joint distribution of $[y_1, y_2]$

---

3 The subscript “$x^1 \rightarrow x^0$” indicates that the attributes data from period $t_1$ will be “replaced” by data from period $t_0$. 
\[ f(y_1, y_2 | T = t_1) = f(y_1, y_2; \sigma_{\mu_1}, \theta_1), \]

and

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

Results are shown in Figures 1 and 2, respectively. Notice 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 would remain as in period 0. This can be computed as follows

\[ f_{x^1 \rightarrow x^0}(y_1, y_2) = f(y_1, y_2; \sigma_{\mu_0}, \theta_1). \]

Results are shown in the first panel of Figure 3. Panel A shows the “true” counterfactual with the \( X \) drawn from the period 0 distribution (\( \sigma_C = \sigma_0 \)) and the location preference of 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.
2.3 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 methods developed by DFL. To keep our exposition self-contained, we provide a careful description of the DFL approach using the same notation as in Leibbrandt, Levinsohn, and McCrary (2010).

Bayes’ rule is first used to obtain that
\[ h(x \mid T = t) = P(X = x)P(T = t \mid X = x) / P(T = t) . \]

DFL recognized that
\[
\frac{h(x \mid T = t_0)}{h(x \mid T = t_1)} = \frac{P(T = t_0 \mid X = x)}{P(T = t_1 \mid X = x)} = \frac{1 - P(T = t_0 \mid X = x)}{P(T = t_0)} = \tau_{t_1 \rightarrow t_0}(x) .
\]

One may use expression (5) to substitute \( h(x \mid T = t_0) \) in equation (3) and obtain that

\[
f_{x^1 \rightarrow x^0}(y_1, y_2) = \int f(y_1, y_2 \mid x, T = t_1) h(x \mid T = t_0) \tau_{t_1 \rightarrow t_0}(x) dx .
\]

Notice that this expression differs from equation (1) only by \( \tau_{t_1 \rightarrow t_0}(x) \). DFL refer to \( \tau_{t_1 \rightarrow t_0}(x) \) 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 Leibbrandt, Levinsohn, and McCrary (2010), and 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 \( T_i = t_0 \) to obtain
\[
\hat{P}(T_i = t_0) = N_0 / (N_0 + N_1) .
\]
2) Estimate $P(T = t_0 \mid X = x)$, by estimating a logit model for both periods. The dependent variable equals one if $T_i = t_0$, and explanatory variables include the vector of individual attributes $x_i$.

3) For the subsample of observations where $T_i = t_1$, estimate the predicted values from the logit

$$\hat{P}(T_i = t_0 \mid X = x_i) = \frac{\exp\{x_i \hat{\beta}\}}{1 + \exp\{x_i \hat{\beta}\}}$$

where $\hat{\beta}$ is the parameter vector from the logit regression.

Then, compute the estimated weights as

$$\hat{\tau}_{it_i \rightarrow t_0} = \frac{\hat{P}(T_i = t_0 \mid X = x_i)}{1 - \hat{P}(T_i = t_0 \mid X = x_i) / P(T_i = t_0)}$$

4) For the subsample of observations where $T_i = t_1$, compute the joint density of periods $[y_1, y_2 \mid T = t_1]$ 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 3. It is clear that the reweighting method is able to replicate the counterfactual spatial distribution remarkably well.

2.4 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 $t_1$ cannot be accounted for by changes in individual attributes $x$, we may compute
The second term in 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 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 3, we decompose the changes in $Y$ into those that can be explained by changes in $x$ and those that cannot. The results for the true counterfactual and the DFL estimate are shown in Figure 4. Panels A and C show the portion of the changes in $Y$ accounted for by changes in $x$ in each case (from changes in $\sigma$). Panels B and D show the unexplained portion of the changes in $Y$ that were due to the change in the parameter $\theta$ from the simulation. Again, it is clear from the figure that the DFL counterfactual provides a very good spatial estimate of the true values of these changes.

3. Applications

To illustrate the application of the techniques developed in the previous section, we conduct two analyses of spatial location decisions: 1) the residential location choices of Hispanics in the Washington, DC, and 2) the spatial distribution of new businesses in Quito, Ecuador. For the first analysis, we use data from the 1990 and 2000 US censuses. For the

---

4 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.

5 The data was provided by the National Historic Geographic Information System. It can be found on their website at http://www.nhgis.org.
second, we use administrative data from the Ecuadorian Tax Authority (Servicio de Rentas Internas SRI).

Understanding the location choices of households and firms 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.

3.1 Spatial Distribution of Hispanics in Washington, DC

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 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; and Iceland, 2004, for example.)

Previous studies suggest that demographic characteristics are strongly associated with minority segregation. For example, Bayer, McMillan, and Rueben (2004) use 1990 census microdata in the San Francisco Bay Area to show that sociodemographic characteristics account for much of the observed segregation between 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. This influx of Hispanics has caused a decline in the average level of education and income of Hispanics relative to others in the DC area (see Table 1), which in turn could affect their spatial distribution.

3.1.1 Unconditional spatial distribution

To compute the distribution of residential location choices, ideally, we would like to know the exact location (and demographic characteristics) of every Hispanic individual living in DC. Those data are not publicly available. The US Census, however, provides aggregate population counts for each census block, block group and tract in the nation. In the DC CBSA, there were 34,795 blocks in the 1990 census with a total population of 4.1 million and 38,982 blocks in the 2000 census with a total population of 4.8 million. To compute the spatial
distribution of Hispanics across DC, we will assume that the residences of all Hispanics living in a particular census block are located at the block's centroid. Given the large number of blocks in the DC area, this assumption should have little effect on our estimates of the spatial distribution.

The block data are used to compute the spatial distribution of Hispanics in 1990 and 2000 and results are shown in Figures 5 and 6, respectively. The spatial distributions were estimated using a product kernel with a 500x500 grid of points. Figure 7 displays the change in the density of the Hispanic population that lives at each grid point. Because the 3-D surface plots can be difficult to interpret at a metropolitan level, Figure 8 shows the same information as Figure 7 in a 2-D surface map of DC. These figures show that there was a decline in the density of Hispanics living in the center of the city and smaller decreases in the Arlington, Virginia and Silver Spring, Maryland areas. These decreases were offset by increases in the Hispanic density in north DC, Rockville, Wheaton, and Hyattsville, Maryland, and areas of Arlington, Virginia. Notice that Figures 7 and 8 provide information about the changes in the density of Hispanics living in a specific location between 1990 and 2000; they do not reflect changes in the total number (level) of Hispanics living in that location.6

The spatial distribution of Hispanics can also be estimated using block group data. As it was the case with the block data, we are forced to make the assumption that all Hispanics living in a census block group are located at its centroid. Because census block groups are significantly larger than blocks, one may worry that this assumption may affect the estimates of the spatial distribution. Figure 9 shows the change in the spatial distribution of Hispanics between the 1990 and 2000 census using the block group level data. Due to the difference in level of aggregation, Figures 8 and 9 are not identical, but they are notably similar. It seems that census blocks in the

---

6 Between 1990 and 2000, the Hispanic population in DC grew from 220,928 to 390,574.
DC area are small enough to provide an accurate representation of Hispanics' location patterns.7

Between 1990 and 2000 there have been significant changes in the spatial distribution of Hispanics in the DC area. During the same period, the characteristics of the Hispanic population in DC also changed. Table 1 contains summary statistics of population counts, age, education, and income levels by race and ethnicity. Between 1990 and 2000, the Hispanic population in DC grew from 220,928 to 390,574, a 77% absolute increase and an increase from 5% to 8% of the population of DC. However, the income level of Hispanics relative to the DC CBSA average declined from 62% to 54% over the same period. The 2000 DC Hispanic population was also younger (more 0-17 year olds) and less educated (the proportion of adults with no high school diploma increased). How do changes in these underlying characteristics of age, education, 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 2000 would look if the distribution of demographic characteristics would have remained constant as in 1990.

3.1.2 Counterfactual spatial distribution

Before we can apply the DFL reweighting method, we need to estimate the demographic characteristics and location of every Hispanic in DC. As it was discussed before, the census block and block group data do not provide the characteristics of each individual. Instead only aggregate demographic characteristics are reported. For example, the census reports population counts for each block by age group, race and ethnicity. Block group data contain additional

7 For the analysis of block groups, we only include those with over 25 individuals. With this restriction, there were 2,677 block groups with 4.1 million inhabitants in the 1990 census and 2,939 block groups with 4.8 million inhabitants in the 2000 census.
information about education levels and average income. 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 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, and median education level is a high school diploma, 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. For the sake of completeness and comparison, the estimation is separately performed using block and block group data.

The counterfactual spatial distribution with block level data is estimated using a single covariate: age. In each block in both periods, vector $x$ includes only the median age and the median age taken to the second, third, fourth, and fifth powers. We then estimate the spatial counterfactual to understand how the 2000 distribution of Hispanic would have looked with the age breakdown of Hispanics in the 1990 census. From Equation (7), we decompose the changes in the spatial distribution of Hispanics into those explained by the changes in $x$ (Figure 10, Panel A) and those that are not (Figure 10, Panel B). From the two panels, we can see that changes in Hispanic age distribution account for some but not much of the changes in the spatial distribution of Hispanics in DC.

With block group data, a richer set of $x$ variables can be included in the DFL reweighting at the cost of including fewer “representative” Hispanics due to the greater aggregation of the block group data. For the block group DFL, we include the Hispanic age breakdown from the block level analysis and add dummies for the median level of education (high school, college, and graduate degrees) and per capita income relative to the CBSA-wide per capita income. For
the per capita income variable, we normalize the income of Hispanics at the block group level by the average income of all individuals in the DC CBSA in each census year. Again, we include each non-dummy $x$ variable taken to the first through fifth powers in the DFL logit regression. Figure 11 Panel A shows the portion of the change in the spatial distribution of Hispanics that can be explained by changes in the distribution of age, education, and relative income. Panel B shows the unexplained portion of the change that was due to other factors. With the inclusion of more $x$ characteristics, the block group level DFL accounts for more of the change in the spatial distribution of Hispanics than the block level data (Figure 10). However as Figure 11, Panel B shows, a large amount of the change is still not accounted for by the changes in the underlying characteristics of Hispanics.8

3.2 Spatial Distribution of New Businesses in Quito, Ecuador

In this section, we use our methodology to evaluate how much changes in the demographic characteristics of new firms' shareholders account for changes in the spatial distribution of new firms in Quito, Ecuador. The location choices of firms and its implications for firms' productivity have been extensively studied in the literature. Spatial concentration of firms generates agglomeration economies offering productivity advantages for firms located within a cluster. A large literature has analyzed and measured the sources and the extent of these agglomeration economies (for an extensive review, see Glaeser and Gottlieb, 2009; Puga, 2010; and Rosenthal and Strange 2001 and 2003) and, presumably, changes in the spatial distribution of firms over time will have implications for overall firms' productivity and the extent of

---

8 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.
agglomeration. The decomposition exercise we propose allows researchers to assess what are the determinants of the shift of the spatial distribution over time.

### 3.2.1 Unconditional spatial distribution

In this application, we explore how changes in the characteristics of business owners are related to changes in the spatial distribution of new businesses. In this case, we look at new businesses formed in 2001 and 2002 and compare them to businesses formed in 2008 and 2009. We focus on the canton of Quito as it is the one with the largest number of new business formations in Ecuador in the periods analyzed. \(^9\)

We use administrative data provided by the Ecuadorian Tax Authority (SRI) on the date of incorporation of all formal-sector firms in Ecuador between 1998 and 2012. These data include business characteristics and date of incorporation. SRI also shared administrative data containing the individual characteristics of each individual shareholder in the sample. We match the shareholder data with the business data using a unique individual tax ID. To determine the firm’s exact location, we use information from the (restricted version) of the 2010 Economic Ecuadorian Census (EEC). \(^10\) In the 2001-02 period, 562 new businesses were formed in Quito, and in 2008-09, 823 new businesses were formed. \(^11\)

Figure 12 shows the spatial distribution of new businesses in Quito in 2001-02 (Panel A) and 2008-2009 (Panel B). New businesses are assigned to the administrative zone in which they

---

\(^9\) The canton is the largest political subdivision at the provincial level. The canton of Quito includes the metropolitan area of Quito as well as some of the rural surrounding area. It contained approximately 1.84 million people in the 2001 census.

\(^10\) In 2010, all economic establishments in Ecuador were asked to fill a questionnaire (the EEC), where among many other variables, its location was recorded. Location of the firm was not exactly determined (i.e. no latitudes and longitudes were recorded). Instead, the census block where the firm operates was identified.

\(^11\) Notice that the number of new firms in our sample are those that were part of the EEC in 2010; that is, firms that were economically active in 2010 for which we have ownership information.
were formed. These zones are similar in size to US census block groups. From the figure, we can see that the north-central portion of the city experienced the largest share of new business formation in both periods. However as Panel C shows, the city also experienced a change in the spatial distribution between the two periods as well.

Table 2 includes information about the demographic characteristics of new firm owners in the two periods. The characteristics of firm owners changed substantially between the two periods. The share of male shareholders declined by nearly 6% from 70.6% to 64.7%. The educational distribution of shareholders also changed. A larger share of new firm owners in 2008-09 have secondary or lower levels of education. In 2001-02, an average of 54.8% of new business owners had a college or graduate education. That declined to 46.3% in 2008-09. The age distribution of average shareholders also changed between the two periods, as individuals 40 and under comprised a larger average percentage of new business shareholders in 2001-02 than in 2008-09. We now explore how these changes in the averages characteristics of new firm owners are associated with changes in the spatial distribution of new firms.

3.2.2 Counterfactual spatial distribution

Figure 12 Panel D shows the DFL counterfactual distribution for the 2008-09 period reweighted to match the distribution of shareholder characteristics from 2001-02. By comparing the counterfactual to the true distributions, we can see that the two are similar. However, Panels E and F show the change in the spatial distribution that is explained and unexplained by the counterfactual respectively. The direction of change predicted by the counterfactual generally corresponds to the change observed. However, the share of the change explained by the counterfactual is generally smaller in magnitude than the share unexplained. Therefore, the
changes in the average observed characteristics of new-business shareholders can account for some, but not the majority, of the change in the spatial distribution of business formation in Quito, Ecuador between 2001 and 2008.

4. Conclusion

In this paper, we extended the notion of counterfactual distributions to spatial problems. We apply the DiNardo, Fortin, and Lemieux 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 phenomena: 1) how individual characteristics have affected the location decisions of Hispanics in the Washington, DC area using block and block group data from the 1990 and 2000 US censuses, and 2) how the change in the average characteristics of shareholders affected the spatial distribution of new firms in Quito, Ecuador between 2001 and 2008.

We hope our approach is useful in other applications.
References


Tables

Table 1: Summary Statistics by Race from DC CBSA Block Groups

<table>
<thead>
<tr>
<th></th>
<th>Whites</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population (Percent of CBSA)</td>
<td>2,635,225 (64%)</td>
<td>2,651,112 (55%)</td>
<td>1,050,514 (25%)</td>
<td>1,249,666 (26%)</td>
<td>220,928 (5%)</td>
<td>390,574 (8%)</td>
</tr>
<tr>
<td>Per Capita Income (Percent of CBSA Average)</td>
<td>$24,629 (117%)</td>
<td>$37,616 (123%)</td>
<td>$14,062 (67%)</td>
<td>$21,380 (70%)</td>
<td>$13,007 (62%)</td>
<td>$16,630 (54%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age (% in each range)</th>
<th>Whites</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-17</td>
<td>22%</td>
<td>23%</td>
<td>27%</td>
<td>29%</td>
<td>27%</td>
<td>30%</td>
</tr>
<tr>
<td>18-24</td>
<td>10%</td>
<td>8%</td>
<td>12%</td>
<td>9%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>25-54</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>55-64</td>
<td>8%</td>
<td>10%</td>
<td>7%</td>
<td>8%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>&gt;64</td>
<td>9%</td>
<td>11%</td>
<td>8%</td>
<td>8%</td>
<td>4%</td>
<td>3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education* (% in each group)</th>
<th>Whites</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No HS Diploma</td>
<td>11%</td>
<td>8%</td>
<td>26%</td>
<td>19%</td>
<td>36%</td>
<td>43%</td>
</tr>
<tr>
<td>HS Diploma</td>
<td>46%</td>
<td>42%</td>
<td>55%</td>
<td>58%</td>
<td>41%</td>
<td>37%</td>
</tr>
<tr>
<td>College Degree</td>
<td>24%</td>
<td>27%</td>
<td>12%</td>
<td>15%</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>19%</td>
<td>23%</td>
<td>7%</td>
<td>9%</td>
<td>10%</td>
<td>9%</td>
</tr>
</tbody>
</table>

* Educational attainment for those over 25

Source: 1990 and 2000 US Census data from the National Historic Geographic Information System. It can be found on their website at http://www.nhgis.org. The summary statistics were compiled from block group level data.
<table>
<thead>
<tr>
<th></th>
<th>2001-02</th>
<th>2008-09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Businesses</td>
<td>562</td>
<td>823</td>
</tr>
<tr>
<td>Male</td>
<td>70.6%</td>
<td>64.7%</td>
</tr>
</tbody>
</table>

**Education**

<table>
<thead>
<tr>
<th>Education</th>
<th>2001-02</th>
<th>2008-09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than Primary</td>
<td>0.3%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Primary</td>
<td>4.1%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Secondary</td>
<td>40.9%</td>
<td>47.8%</td>
</tr>
<tr>
<td>Tertiary or Above</td>
<td>54.8%</td>
<td>46.3%</td>
</tr>
</tbody>
</table>

**Age**

<table>
<thead>
<tr>
<th>Age</th>
<th>2001-02</th>
<th>2008-09</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤30</td>
<td>30.2%</td>
<td>24.1%</td>
</tr>
<tr>
<td>31-40</td>
<td>32.0%</td>
<td>32.2%</td>
</tr>
<tr>
<td>41-50</td>
<td>22.6%</td>
<td>23.4%</td>
</tr>
<tr>
<td>51-60</td>
<td>10.1</td>
<td>13.3%</td>
</tr>
<tr>
<td>&gt;61</td>
<td>5.0%</td>
<td>6.9%</td>
</tr>
</tbody>
</table>
Figures

Figure 1: Parametric Example – Period 0 Spatial Distribution ($\theta_0 = 0.2, \sigma_0 = 0.5$)

Figure 2: Parametric Example – Period 1 Spatial Distribution ($\theta_1 = 0.5, \sigma_1 = 1.0$)
Figure 3: Counterfactual Spatial Distribution

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

B. Estimated Counterfactual
Figure 4: Counterfactual Spatial Distribution – Changes in Distribution Explained and Unexplained by Changes in $X$

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 $X$
Figure 5: 1990 Distribution of Hispanics with Block Data

Figure 6: 2000 Distribution of Hispanics with Block Data
Figure 7: Change in the Spatial Distribution of Hispanics with Block Data (2000 – 1990)

Figure 8: Map of DC CBSA – Change in the Spatial Distribution of Hispanics with Block Data (2000 – 1990)
Figure 9: Map of DC CBSA – Change in the Spatial Distribution of Hispanics with Block Group Data (2000 – 1990)
Figure 10: Explained and Unexplained Change in the Spatial Distribution of Hispanics with Block Data

Figure 11: Explained and Unexplained Change in the Spatial Distribution of Hispanics with Block Group Data


Figure 12: Spatial Distribution of New Firms in Quito, Ecuador

A. New Businesses in 2001-02

B. New Businesses in 2008-09
Appendix 1. Parametric Example Model

In this section, we provide details about the parametric example in Section 2.2. The location coordinates chosen by individual $i$ in period $j$ are given by $y_{1i}^j = \epsilon_{1i}^j X_{1i}^j$ and $y_{2i}^j = \epsilon_{2i}^j X_{2i}^j$, where time period $j = \{0, 1\}$. We assume that $X$ is a random variable and that its distribution depends on a sole parameter $\sigma_{ij}$. In particular, we let $X_i^j = \beta_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij})$, and $\mu_j \sim \mathcal{N}(0, \sigma_{\mu j})$.

We assume that $\epsilon_1^j$ and $\epsilon_2^j$ are random variables independent of $X$. For instance, we let $\epsilon_k^j = \pi^j W_k + (1 - \pi^j) Z_k$, $W_k \sim \mathcal{N}(\mu_{Wk}, \sigma_{Wk})$, $Z_k \sim \mathcal{N}(\mu_{Zk}, \sigma_{Zk})$, $k = \{1, 2\}$, and $\pi^j \sim \text{Bernoulli}(\theta_j)$. In this simple model, the only parameters that are time dependent are $\sigma_{\mu j}$ and $\theta_j$. Notice that $\sigma_{\mu j}$ measures the dispersion of the covariates in each time period, while $\theta_j$ 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, 10,000) 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 equal to $\mu_{W1} = -2$, $\sigma_{W1} = 1$, $\mu_{Z1} = 2$, $\sigma_{Z1} = 1$, $\mu_{W2} = 2$, $\sigma_{W2} = 1$, $\mu_{Z2} = -2$, and $\sigma_{Z2} = 1$. 

36