December 2018

THREE ESSAYS IN URBAN AND REGIONAL ECONOMICS

Boqian Jiang
Syracuse University

ABSTRACT

This dissertation comprises three chapters that are related to the research topics in Urban and Regional Economics. The first chapter examines whether economic self-interest associated with homeownership motivates homeowners to vote more than renters in U.S. local elections. To control for the self-selection of homeownership, I use national election turnout as the counterfactual outcome. Since policy discussions in national elections are targeted more at the national level, the disparity in political participation between homeowners and renters should be reduced. Results based on election data from three U.S. cities confirm these hypothesis, which suggest that local policies may tend to cater to the tastes of homeowners over renters. The second chapter develops a new method to identify and control for selection when estimating the productivity effects of city size. For single peaked factor return distributions, selecting out low-performing agents has limited effect on modal productivity but reduces the CDF evaluated at the mode. Spillovers from agglomeration have the reverse effect. Estimates based on law firm productivity, wages for married women and wages for full-time men all confirm that selection contributes to urban productivity and that doubling city size causes productivity to increase by 1-2.5 percent. The last chapter uses border discontinuity design to study the long-run effect of British colonial rule on the state building in Africa. British colonial legacy is featured with ethnic segregation and stronger executive constraints, which may have undermined state centralisation. Using micro-data from anglophone and francophone countries in sub-Saharan Africa, we find that anglophone citizens are less likely to identify themselves in national terms (relative to ethnic terms). Evidence on taxation, security and the power of chiefs also suggests weaker state capacity in anglophone countries. These results highlight the legacy of colonial rule on state-building.

THREE ESSAYS IN URBAN AND REGIONAL ECONOMICS

by

Boqian Jiang

B.A., Southwestern University of Finance and Economics, 2011
M.A., Toulouse School of Economics, 2013

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in *Economics*.

Syracuse University
December 2018

ACKNOWLEDGMENTS

First and foremost, I offer my sincerest gratitude to my advisor Stuart Rosenthal who has supported me throughout my Ph.D. with his knowledge and patience. I am deeply in debt for his trust, encouragement, and guidance. I am also very grateful for the valuable discussions and career advice from Hugo Jales, Alfonso Flores-Lagunes, and Abdulaziz Shifa.

I am also grateful to Badi H. Baltagi, William C. Horrace, Yilin Hou, Jeffrey D. Kubik and other department seminar participants for their helpful comments and suggestions on my papers.

Lastly, I want to express special appreciation to my parents – Juan Xie and Zhiming Jiang. I would not have gone this far without their unconditional love. I also would like to thank my friends, especially Jerome Flaviyan, Ying Lei Toh and Giridaran Subramaniam, for their continuing support during my Ph.D. and the job market season.

Syracuse, New York
April 2018

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1: Homeownership and U.S. Local Election Turnout

Chapter 2: Separating Selection From Spillover Effects: Using the Mode to Estimate the Return to City Size

Chapter 3: Colonial Legacy, State-building and the Salience of Ethnicity in Sub-Saharan Africa

LIST OF FIGURES

Chapter 2: Homeownership and U.S. Local Election Turnout

Chapter 3: Colonial Legacy, State-building and the Salience of Ethnicity in Sub-Saharan Africa

Chapter 1.

Homeownership and U.S. Local Election Turnout

Boqian Jiang
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, New York, 13244-1020
bjiang03@syr.edu

Abstract

This paper provides evidence that economic self-interest associated with homeownership affects voter turnout in local elections in the United States. Compared to renters, homeowners are financially invested in their communities and are less mobile. Therefore, homeowners should care more about local policies and have incentives to engage actively in local politics. The disparity in political participation between homeowners and renters, however, should diminish in presidential elections for which policy discussions are more targeted at the national-level. These hypotheses are tested using block-level election panel data. Fixed effects models and a control function approach are used to identify the effect of homeownership on voter turnout in off-year mayoral elections relative to presidential elections. Results show that mayoral election voter turnout increases with the local homeownership rate. This suggests that local policies may tend to cater to the tastes of homeowners over renters.

JEL Classification: R2, R3, D72

Key words: homeownership, voter turnout, local election.

1. Introduction

Local governments in the United States are the primary provider of local public goods and enact zoning laws that affect allowable patterns of land use. Collectively, local government expenditures also account for one-tenth of U.S. GDP and local governments collect as much in tax revenue as the federal government (Oliver et al., 2012). Nevertheless, compared to national-level elections, U.S. local races often have very low voter turnout. Table 1 summarizes mayoral and presidential election turnout rates from Philadelphia, Seattle and Chicago over the period between 2003 and 2013 (with turnout rates measured at the census block group level).[1] While presidential election turnout averages roughly 60%, mayoral election turnout is much lower, ranging from 20% to 40%. Turnout rate variation across census blocks – measured by standard deviation divided by the mean – is also higher in mayoral races. Given local government's consequential role, the limited turnout in local elections is a source of concern. The reason is that voter turnout could affect how local governments enact policies and whether local policies are representative of the electorate (Hajnal and Trounstine, 2005; 2010).

The question of why voter turnout is so low in local elections is related to the literature on the homevoter hypothesis (Fischel, 2001; Brunner et al., 2001; Brunner and Sonstelie, 2003; Dehring et al., 2008; Hilber and Mayer, 2009; Ahlfeldt, 2011; Ahlfeldt and Maennig, 2015). Previous studies by Fischel (2001) and others have examined how capitalization effects associated with local policies may affect home-voters preferences over local policy initiatives as compared to lease-voters. That literature, however, has largely overlooked related effects on local election turnout. This paper fills that gap by providing evidence that voter turnout in local elections is driven in part by voter economic self-interest related to homeownership status.

[1] For each city, the statistics are organized into three election cycles and the election cycles are four years apart.

For two reasons, homeowners have stronger economic incentives to vote in mayoral elections relative to renters. From an investment perspective, the value of a homeowner's house – the largest investment for most U.S. households[2] – is tied to local fiscal services and amenities provided by the municipal government (Rosen, 1974; Ross and Yinger, 1999; Yinger, 2015). From a consumption perspective, homeowners are also less mobile and hence receive longer utility flows from local public goods. Renters, in contrast, are less financially invested in their communities and more mobile (Rosenthal, 1988; Ioannides and Kan, 1996). Therefore, renters are less likely to internalize the long run effect of their local political decisions and have less incentive to vote in local elections.

The empirical challenge in this paper is to identify the causal effects of homeownership on mayoral election voter turnout. It has been widely recognized that homeownership and political participation may be endogenously correlated (DiPasquale and Glaeser, 1999; Keyssar, 2009; Engelhardt et al., 2010). Failing to control for unobserved confounders may bias estimates of the effect of the owner-occupancy rate on voter turnout.[3] To address the endogeneity concern, I use two strategies to control for census block level unobservables, a block-level fixed effects model and a control function approach. Both models use presidential elections as the counterfactual.

National-level elections and related policy initiatives are by definition less focused on local issues, the provision of local public goods, and local property values. For that reason, and drawing

[2] According to a report by The Federal Reserve Board, even at the bottom of 2008 housing crisis, housing wealth still counts as one-half of the total household net wealth in the United States. For the median household, housing wealth counts for almost two-thirds of their total wealth. Link: https://www.federalreserve.gov/pubs/ifdp/2011/1027/ifdp1027.htm

[3] There has been a few attempts in the literature to address the endogeneity. For instance, DiPasquale and Glaeser (1999) instruments individual homeownership using group-average homeownership rates from the corresponding socio-demographic groups. However, their instrument is far from perfect since individual unobservables may correlate with membership to a socio-demographic group. Engelhardt et al. (2010) exploits the random assignment of home-purchase subsidies to low-income renters in a field experiment conducted in Tulsa, Oklahoma. Despite a cleaner study design, their sample is not representative of the general population and the sample size is limited.

on capitalization arguments from the homevoter literature, homeowners and renters should display more similar tendencies to vote in national elections relative to local elections, all else equal. Hence in a well-specified model, block-level homeownership rates should not be correlated with presidential election turnout provided one sufficiently controls for socioeconomic differences between homeowners and renters.

In the fixed effects models to follow, I assume that the census block confounders are time-invariant. After differencing away time-invariant unobserved confounders, the homeownership rate strongly affects mayoral election turnout but does not affect presidential election turnout. The sharp differences in estimates for mayoral and presidential elections provide evidence that economic incentives contribute to voter turnout and motivate homeowners to be more likely to vote in mayoral elections than renters.[4]

One may argue against the assumption that the block-level confounders are time-invariant. Relaxing this assumption motivates my second identification strategy – the control function (CF) approach. I directly model the time-varying confounders in mayoral regressions as a function of presidential election turnout rate. To obtain identification, the CF approach imposes other moderate assumptions that are clarified later in the paper. Empirically the two models deliver very similar estimates.

An interpretation of my identification strategies is that, by controlling for block-level unobservables, both models indirectly absorb individual-level confounders that contribute to residential sorting. According to the Tiebout sorting theory, households may sort into

[4] That interpretation is further strengthened by including household mobility contorls (principally census-block residential turnover rates) into the model to help separate investment and consumption motives for voter turnout. Several other time-varying census block level socioeconomic attributes are also taken into account, including income, education, age distribution and share of households that are married. As noted above, the most robust models include block-level fixed effects and identify off of within-block temporal variation in the data.

neighborhoods according to their needs and willingness to pay for local public amenities (Tiebout, 1956). In a sorting equilibrium, homeowners and renters living in the same neighborhood share lots of similar characteristics. Therefore, census block-level unobserved attributes may correlate with the individual characteristics that cause the endogenous correlation between homeownership and voter turnout.[5]

To conduct the analysis, I assembled a novel election panel data for Philadelphia, Chicago, and Seattle over the period between 2002 and 2013. My key specification shows that when a census block switches from fully rental into fully owner-occupied, its mayoral election turnout rate will increase by approximately five percentage points, which is equivalent to a 23 percent increase compared to the mean. The tenure composition change does not affect presidential election turnout.

Results from this paper suggest that renters are under-represented in U.S. local elections. As renters participate less in local races due to insufficient economic self-interest, local politicians may design policies to please the high turnout group – homeowners – to gain electoral support.[6] Such favoritism may lead to policies protecting property value appreciation (e.g. strict zoning laws). Glaeser et al. (2005) points out that change in land regulation regime explains the scarcity of house development in the most expensive U.S. housing market. Ortalo-Magn and Prat (2014) theorizes how homeowners affect urban growth control through the local political process. Total social welfare may also be impaired by the tightening housing supply as it impedes an efficient spatial allocation of labor (Hsieh and Moretti, 2015).

[5] Indeed, Minkoff (2014) finds that the quality of city-provided public goods in a community is highly correlated with residents' tendency to vote in New York City.

[6] In San Francisco, a city with a roughly thirty-five percent owner-occupancy rate, households organize into hundreds of politically powerful neighborhood groups (e.g., Telegraph Hill Dwellers) to promote policies limiting new house development. See *Kim-Mai Cutler*, "How Burrowing Owls Lead To Vomiting Anarchists (Or SF's Housing Crisis Explained)", TechCrunch, April 2014. Link: http://techcrunch.com/2014/04/14/sf-housing/

The remainder of the paper is organized as follows. Section 2 gives a detailed discussion of the empirical specializations and identification strategies. Section 3 provides a description of the data source and summary statistics. Section 4 presents the empirical results and robustness check. Then the paper ends with some concluding remarks in Section 5.

2. Empirical Specification and Identification

The basic empirical specification is given as follows:

$$Mayoral\ Election\ turnout_{b,c,t} = \alpha_m + \beta_m \times Owner\text{\_occupancy}\ rate_{b,c,t} + \mathbf{X}_{b,c,t}{}'\boldsymbol{\Gamma}_{\mathbf{m}} + u_{b,c,t} \qquad (1)$$

in which b indexes census blocks, c cities and t election cycles. Coefficient β_m measures the impact of owner-occupancy rate on mayoral election voter turnout. This paper attempts to obtain an unbiased estimator of β_m. Vector $\mathbf{X}_{b,c,t}$ represents census block level observables that may correlate with owner-occupancy rate and mayoral election turnout.

The first control to be included in $\mathbf{X}_{b,c,t}$ is the residential mobility measure. As described in the introduction, owning a property makes homeowners have a higher financial stake in the community as well as be less mobile. Staying put means homeowners enjoy a longer utility flow from local public good. This may also motivate them to vote in local elections to affect public goods provision.[7] Controlling for the residential mobility measure aims to separate the effect of investment and consumption motive on voter turnout. However, there is no direct measure of

[7] One may also suspect that moving too frequently may bring hurdle to voter registration and thus depress election turnout. Given the fact that voter registration deadline for most States is 15 – 30 days before the Election day, as long as the majority of the movers do not frequently move immediately before the Election day, the effect of mobility on voter registration should be small.

residential mobility at census block level. I use census block turnover rate – the share of residents moving into the neighborhood within past 12 months – as a proxy for the mobility measure, by assuming that people living in high turnover neighborhoods are in general more mobile. Being aware that the turnover rate is not an adequate mobility measure, I also rely on other covariates to be introduced later to control for residential mobility. In a comprehensive review, Molloy et al. (2011) shows that U.S. residential mobility is closely tied to demographic attributes such as education or age.

The second control is the household median income. According to an analysis done by the real estate database company Zillow[8], owner-occupancy rate almost monotonically increases with household income. Meanwhile, higher income may have a negative impact on voter turnout. Charles and Stephens (2013) shows that higher labor income discourages voter turnout because higher hourly wages implies higher opportunity cost of going to the poll. The third control is the education distribution. Homeownership is highly correlated with education according to a report from First American Financial Corporation.[9] Numerous studies also find that education is a strong driver of election turnout (Sondheimer and Green, 2010; Burden, 2009). Looking at residents 25 years old or above, I generate the share of residents having a high school degree or some college and the share of residents with at least a bachelor's degree. These two variables jointly measure the education distribution of the census block. The fourth control is the share of married household. It has been found that marriage is closely linked to the first-time transition into homeownership (Smits and Mulder, 2008). Marriage may affect turnout through a complex web of channels. For instance, marriage may depress turnout if married couples need to spend time with their children at home, which prevents them from going to the polls. On the other hand, if couples have school-

[8] Link: http://www.zillow.com/research/homeownership-by-income-9419/

[9] Link: http://www.firstam.com/economics/homeownership-progress-index/

aged children that they drop off to school and the polling stations are close to local schools, it might be convenient for them to vote. In this case, marriage may increase voter turnout.[10] The last control is the age distribution. Age is another strong predictor of homeownership and it is well documented in the voting literature that propensity to vote increases with age (Blais, 2000; Wolfinger and Rosenstone, 1980). I use the share of adults between 30 and 60 years old and the share of adults above 60 years old to jointly measure the age distribution of the census block.

After controlling for the observed block-level attributes, standard OLS theory shows that the estimator of β_m is unbiased if owner-occupancy rate and the remaining error term $u_{b,c,t}$ are independent. Nevertheless, $u_{b,c,t}$ may still contain unobserved confounders that are correlated with owner-occupancy rate. I further decompose $u_{b,c,t} = \gamma_m \delta_{b,c,t} + \varepsilon_{b,c,t}$ and define $\delta_{b,c,t}$ as the unobserved block-level confounders. Coefficient γ_m measures the correlation between the confounders and mayoral election turnout. Since $\delta_{b,c,t}$ is unobserved to the researcher, this paper uses two strategies to control for it.

2.1. Fixed Effects Model

The first approach exploits the panel structure of the data. Assuming that the unobserved confounders are time-invariant, I use a fixed effects model to absorb the unobservables. Replacing $\delta_{b,c,t}$ by $\delta_{b,c}$, Equation (1) can be written as:

$$\begin{aligned} Mayoral\ Election\ turnout_{b,c,t} = \quad & \alpha_m + \beta_m \times Owner\text{\_occupancy}\ rate_{b,c,t} \\ & + \mathbf{X}_{b,c,t}{}'\mathbf{\Gamma_m} + \gamma_m \delta_{b,c} + \pi_t + \varepsilon_{b,c,t} \end{aligned} \tag{2}$$

[10] One may argue that it is the existence of school-aged children that matters for married households, since families with children may care more about local public goods provision (such as police, sidewalks, school quality, and so forth.) and therefore more likely to vote. Empirically I find that controlling the share of households with school-aged children instead of the share of married household does not affect the result.

in which I also include election-cycle fixed effects π_t to capture time trend in the data. After the unobservables being absorbed by block-level fixed effects, an unbiased estimator of β_m can be obtained.

The identification assumption in the fixed effects model is that the unobserved confounders are time-invariant. To support this assumption, I use presidential election outcome as a counterfactual. Since national politics are less relevant to property value and local public good provision, the disparate incentive to vote between homeowners and renters should be reduced. Therefore, if the fixed effects can absorb block-level unobserved confounders, homeownership rate should not drive presidential election turnout in a fixed effects model. Replacing the dependent variable in Equation (2) with presidential election turnout, I have:

$$\begin{aligned} Presidential\ Election\ turnout_{b,c,t} = \quad & \alpha_p + \beta_p \times Owner\text{\_occupancy}\ rate_{b,c,t} \\ & + \mathbf{X}_{b,c,t}{}'\mathbf{\Gamma_p} + \gamma_p \delta_{b,c} + \pi_t + \varepsilon_{b,c,t} \end{aligned} \tag{3}$$

In Equation (3), the same time-invariant confounder term $\delta_{b,c}$ may correlate with owner-occupancy rate and thus generates an endogeneous correlation between owner-occupancy rate and presidential election turnout. Coefficient γ_p captures the effect of unobservables on presidential election turnout and it is allowed to be different from γ_m in Equation (2). Coefficient β_p measures the impact of owner-occupancy rate on presidential election turnout. As will be shown in the empirical section, β_p becomes statistically indistinguishable from zero once the fixed effects are introduced.

2.2. Control Function Approach

One may argue against the assumption that the block-level confounders are time-invariant. Relaxing this assumption motivates my second identification strategy – the control function (CF) approach. The basic idea of a CF approach is to control for the unobservables in the error term by modeling them directly (Wooldridge, 2015). Since I can observe turnout rates from both mayoral and presidential elections, I can re-write the confounder term $\delta_{b,c,t}$ in mayoral election regression as a function of presidential election turnout rate. To better explain the CF approach, I rewrite Equation (2) and (3) as:

$$\begin{aligned} Mayoral\ election\ turnout_{b,c,t} = \quad & \alpha_m + \beta_m \times Owner\text{\_occupancy}\ rate_{b,c,t} \\ & + \gamma_m \delta_{b,c,t} + \varepsilon_{b,c,t} \end{aligned} \tag{4}$$

and

$$\begin{aligned} Presidential\ election\ turnout_{b,c,t} = \quad & \alpha_p + \beta_p \times Owner\text{\_occupancy}\ rate_{b,c,t} \\ & + \gamma_p \delta_{b,c,t} + \varepsilon_{b,c,t} \end{aligned} \tag{5}$$

Vector $\mathbf{X}_{b,c,t}$ is dropped for simplicity. The confounder term $\delta_{b,c,t}$ enters both equations and is allowed to have heterogeneous impact across elections (measured by γ_m and γ_p repectively). My coefficient of interest is β_m, which measures the impact of owner-occupancy rate on mayoral election turnout. Notice that neither β_m or β_p can be estimated without bias because $\delta_{b,c,t}$ is unobserved. Rearranging Equation (5), I can write $\delta_{b,c,t}$ as a function of the $Presidential\ election\ turnout_{b,c,t}$. Plugging the new expression of $\delta_{b,c,t}$ into Equation (4), I arrive at:

$$\begin{aligned} Mayoral\ election\ turnout_{b,c,t} = \quad & A + (\beta_m - \frac{\gamma_m}{\gamma_p}\beta_p) \times Owner\_\text{occupancy}\ rate_{b,c,t} \\ & + \frac{\gamma_m}{\gamma_p} Presidential\ election\ turnout_{b,c,t} + \epsilon_{b,c,t} \end{aligned} \qquad (6)$$

in which $A = \alpha_m - \frac{\alpha_p}{\gamma_p}$ and $\epsilon_{b,c,t} = (1 - \frac{\gamma_m}{\gamma_p})\varepsilon_{b,c,t}$. Now all the controls in Equation (6) can be observed from the data. More important, the owner-occupancy rate is independent from the new error term. Therefore, I can obtain an unbiased estimator of the coefficient $\beta_m - \frac{\gamma_m}{\gamma_p}\beta_p$. This is the difference between my coefficient of interest β_m and a scaled version of β_p. Before introducing the identification assumptions, let me briefly discuss the scale parameter $\frac{\gamma_m}{\gamma_p}$. Notice that although $\frac{\gamma_m}{\gamma_p}$ also appears as the coefficient of presidential election turnout rate, it cannot be estimated without bias because the presidential election turnout is correlated with the error term by construction. It is reasonable to presume that the sign of $\frac{\gamma_m}{\gamma_p}$ to be positive since γ_m and γ_p measure the endogenous correlation between owner-occupancy rate and election turnout rates. In the empirical section, I find that the estimate of $\frac{\gamma_m}{\gamma_p}$ is approximately 0.45 and it remains stable in all specifications.

To identify β_m, I need to make one of the two alternative assumptions. One option is assuming $\beta_p = 0$. In this case $\beta_m - \frac{\gamma_m}{\gamma_p}\beta_p = \beta_m$, so that an OLS regression based on Equation (6) returns the unbiased estimate of β_m. The other option is assuming β_p to be non-negative. Under this weaker assumption I can identify the lower bound of β_m given the fact that $\frac{\gamma_m}{\gamma_p}$ is positive. Both assumptions imply that conditional on $\delta_{b,c,t}$ the owner-occupancy rate has a negligible effect on presidential election turnout. Neither of these assumptions is restrictive. According to the fixed effects model estimation results to be shown in Section 4.1, the correlation between owner-occupancy rate and presidential election turnout is not statistically different from zero. Since CF approach further allows $\delta_{b,c,t}$ to be time-varying, this more flexible specification should further push β_p to be close to zero. Therefore, depending on which assumption one is willing to make, either β_m or its lower bound is identified.[11] For the ease of discussion, I assume that $\beta_p = 0$, therefore β_m can be identified from an OLS regression using Equation (6).

3. Data and Summary Statistics

This paper draws upon data from the U.S. Census and election databases from Philadelphia, Chicago, and Seattle. Municipal governments from these three cities digitized their precinct-level election returns from 2002 onward and published them online (see Appendix A for details). The election database reports election type, candidate information, and vote counts for local, state, and federal elections. It also includes records from various idiosyncratic local ballots.

Now I discuss the reasons to choose mayoral and presidential elections as the focus of this study. First, my main identification strategy relies upon using fixed effects to absorb block-level time-invariant confounders. Hence I need to look at elections that were held for multiple times. Second, choosing elections that are common for all three cities allows me to pool data together to improve estimation precision. Presidential elections are a natural candidate for national elections since they are held every four years and are universal for all residents living in the United States. For local elections, I chose mayoral elections because all three cities have the same mayor-council (as opposed to council-manager) governance system that allows city residents to elect their mayors directly. Additionally, all three cities hold their mayoral elections in odd–numbered years (off–

[11] Identification fails only when owner-occupancy rate has a notable negative impact on presidential election turnout. In that case the estimated coefficient is an upper bound of β_m. However, based on the empirical results in Section 4.1 I can rule out this scenario.

year elections) so that the mayoral election turnout is not affected by presidential election turnout (Levonyan, 2013). Another potential candidate for local election is the city-council election. In my data, the city-council and the mayoral election turnout rates are close because they are held at the same time. Voters, given that the marginal cost of participating in city-council elections is zero conditional on already voted in a mayoral race, usually cast votes for both elections. Since mayor tends to be a more viable political player in the mayor-council system (Holbrook and Weinschenk, 2013) and it is the local executive parallel to the U.S. chief executive, this paper focuses on mayoral election results. Other elections such as gubernatorial or congressional elections are not as relevant for local policy as mayoral elections and are therefore not considered in the analysis.

I group the election turnout data into three election cycles according to chronological order. Each election cycle contains a mayoral election and a presidential election that is one year apart. Between 2002 and 2015, Philadelphia and Chicago held mayoral elections in 2003, 2007 and 2011; Seattle held mayoral elections in 2005, 2009 and 2013. Correspondingly, there are three presidential elections in 2004, 2008 and 2012. The first three columns in Table 2 summarizes the election grouping. For instance, the first election cycle includes 2003 Philly, 2003 Chicago, and 2005 Seattle mayoral elections as well as the 2004 presidential election from these three cities.

To merge the election data with the Census data, I map the raw precinct-level voting records into census blocks (2010 definition) using a crosswalk from the Missouri Census Data Center.[12] Census blocks are the smallest geographic unit for which the Census publishes its data. A census block typically has a population between 600 and 3,000. The average census block population size is 1,000 for the three cities. Voting precincts are slightly smaller than census blocks

[12] http://mcdc.missouri.edu/websas/geocorr12.html

geographically. In Philadelphia, 1,686 precincts are mapped into 1,333 blocks. In Chicago, 2,571 precincts are mapped into 2,171 blocks and, in Seattle, 960 precincts are mapped into 479 blocks.

Census data at the block level is accessible from two Census data products: the Decennial Census and the American Community Survey (ACS) 5-year-estimates. The main difference between them is that the Decennial Census provides data gathered at a "point of time" while ACS produces estimates using data gathered within a "period of time". Since socio-demographic attributes of small geographic areas are likely to remain constant within a 5-year time window, ACS 5-year-estimates are designed to provide estimates describing the average attributes of an area over the corresponding time period.[13] Moreover, Decennial Census data is available once per decade while ACS 5–year–estimates have more frequent time coverage.[14] The data merging between election panel and Census data is summarized in Table 2 Column (3). There is no corresponding Census data coverage for elections held during the first election cycle (2003 to 2005). Hence I approximate the block-level attributes during this period by taking the average between 2000 Decennial Census data and ACS 2005 – 2009 5–year–estimates.[15] For election data from the second and the third election cycles, I merge them with Census data from ACS 2006 - 2010 5-year-estimates and ACS 2010 - 2014 5-year-estimates, respectively.[16]

[13] For more information of ACS data, please look at the "A Compass for Understanding and Using American Community Survey Data" document prepared by Census. http://www.psc.isr.umich.edu/dis/acs/handouts/Compass_Appendix.pdf

[14] Currently ACS has 6 rounds of 5–year–estimates available: 2005 – 2009, 2006 – 2010, 2007 – 2011, 2008 – 2012, 2009 – 2013, 2010 – 2014.

[15] The approximation can be done for all variables except the neighborhood turnover rate, which measures share of residents moved into the neighborhood within past 12 months. It is available in ACS data but not in 2000 Decennial Census, which only provides the share of residents moved into the neighborhood within past five years. Neighborhood turnover rate appears to have a slightly downward trend over the years. It is mean value decreases from 0.165 in 2005 - 2009 round data to 0.158 in 2010-2014 round data. One option for obtaining approximation is extrapolating the original sequence by assuming linearity. Since the estimation will include election cycle fixed effects which can take care any form of linear/non-linear trend in variables, this paper directly uses 2005 - 2009 data to proximate the neighborhood turnover rate in first election cycle.

[16] All the data merging is based on the 2010 geographic definitions of census blocks. However, 2000 Decennial Census and ACS 2005-2009 5-year-estimates files are coded using 2000 census blocks boundaries while other years ACS data are coded using 2010 census blocks boundaries. The Census Bureau does not provide correspondence file

The block-level voter turnout rate is measured by the total number of votes over the size of voting age population (United States citizens 18 years of age or older) in the block.[17] Due to sampling error in voting age population estimation, there are some blocks with turnout rates larger than one. The empirical analysis to follow, I only use census block samples in which the turnout rate is never larger than one. For instance, if block A has a turnout rate larger than one in election cycle three's presidential election, all records from block A are dropped. After dropping those census blocks I obtain a balanced election panel in which each election cycle is composed of 1,129 Philadelphia census blocks, 1,874 Chicago census blocks and 377 Seattle census blocks. In the robustness checks, I show that all results hold using the original full sample, as well as to as when restricting the sample to blocks with turnout rates between 5% and 95%.

Table 3 provides mean, median and standard deviation for homeownership rate and turnout in mayoral and presidential elections. The first thing to notice is that in each of the cells the mean and median are close to each other, indicating that the turnout rates do not have a skewed distribution. Compared to presidential election turnout, mayoral election turnout is lower.

4. Empirical Results

4.1. Fixed Effects Model

The empirical evidence from the fixed effects model is presented in three parts. First, I present estimation results from Equation (2) and (3). This allows me to compare the effect of owner-occupancy rate on mayoral elections ($\hat{\beta}_m$) versus on presidential elections ($\hat{\beta}_p$). Second, I

to link the 2000 census blocks to the 2010 census blocks. Therefore, I build correspondence file between 2000 and 2010 census blocks using GIS software (see Appendix B for detailed GIS work description).

[17] Voting age population estimates are obtained from a separate tabulation from 2000 Decennial Census and the "Voting Age Population by Citizenship and Race (CVAP)" estimates from the ACS 5-year-estimates. See Appendix C for details.

statistically test the coefficient difference between $\hat{\beta}_m$ and $\hat{\beta}_p$ by pooling mayoral and presidential elections data together to run regression with interaction terms. Last, I present further evidence of obtaining identification using the fixed effects model.

Table 4 presents estimation results based on Equation (2) which uses mayoral election turnout rate as the dependent variable. Robust standard errors are clustered at the census block level to account for time series correlation. The first column shows the raw correlation between owner-occupancy rate and mayoral election turnout. City fixed effects are added to account for city-level heterogeneity. The coefficient on the owner-occupancy rate is positive and significant. It indicates that when a block switches from fully rental into fully owner-occupied, its mayoral election turnout rate increases by 19.7 percentage points, which is almost a 70 percent increase relative to the mean.

Column (2) adds in the mobility control. Not surprisingly, controlling for neighborhood turnover rate drives down the owner-occupancy rate coefficient because it separates the two channels through which homeownership affects voter turnout. Homeowners may be more likely to vote locally because: (a) they have higher financial stakes in the community, (b) they are less mobile so they enjoy a longer utility flow from local public goods. With the mobility control, the coefficient on owner-occupancy rate declines from 0.197 to 0.158. The negative coefficient on the mobility control is consistent with the idea that mobile residents have less incentive to participate in local elections.

Column (3) substitutes city fixed effects with census block fixed effects. Fixed effects absorb time-invariant census block unobserved confounders that may cause an endogenous correlation between homeownership rate and voter turnout. In the census block fixed effects model, coefficient estimates on owner-occupancy rate and mobility control both decrease. Finally, Column (4) brings the full set of census block controls. Owner-occupancy rate still has a notable impact on mayoral election turnout. Holding all else equal, if a block a block switches from fully rental into fully owner-occupied, its mayoral turnout rate increases by 4.7 percentage points. This is equivalent to a 23 percent increase compared to the mean. The coefficient on the mobility control dropped considerably in Column (4). This is unsurprising because mobility is highly correlated with demographic controls such as age and marital status (Molloy et al., 2011). Therefore, variation in the mobility control is partially captured by other variables. The coefficient on household median income confirms the results in Charles and Stephens (2013). Census blocks with higher median household income have lower turnout rate. The coefficients on education distribution show that there is a non-monotonic relationship between education attainment and voter turnout. Compared to residents with less than a high school degree, the share of residents with high school or some college is negatively correlated with turnout while the share of residents with at least a college degree positively correlates with voter turnout. The share of married households does not seem to have a significant correlation with mayoral election turnout. The age distribution coefficients indicate that voter turnout is positively correlated with voter age. Adults above 60 years old have strongest tendency to vote among all groups.

Columns (5) - (8) follow the same specifications as in Columns (1) - (4) but use the log of mayoral election turnout rate as the dependent variable. Transformation into logarithms serves two purposes. First, it accounts for potential non-linearity between voter turnout and the control variables. Second, it allows for the slope coefficients to be interpreted as percentage changes in the turnout rate. This helps to compare mayoral election results to the presidential ones because the two elections have different turnout averages. A similar pattern repeats in the non-linear model.

The owner-occupancy rate remain remains positive and significant in all columns. And the estimated coefficient on the mobility control becomes noisy as more controls are added.

Having established that owner-occupancy rate drives mayoral elections turnout in a fixed effects model, I move to the benchmark presidential election results in Table 5. The first column of Table 5 presents the raw correlation between owner-occupancy rate and presidential election turnout rate. According to the coefficient estimate, when a block switches from fully rental into fully owner-occupied, its presidential election turnout increases by 9.6 percentage points. This is a 17.5 percent increase compared to the average presidential turnout rate. Adding in mobility controls in Column (2) does not change the owner-occupancy rate coefficient by much. In Column (3), replacing city fixed effects with census block fixed effects erases the previously significant coefficient on owner-occupancy rate. Block fixed effects capture the endogenous correlation between homeownership and presidential election turnout. The fixed effects also reduce the coefficient on the mobility control. People's tendency to vote in the presidential election should not be affected by how mobile they are since national policies affect the well-being of all residents in the country. Including the full set of control in Column (4) does not result in any further changes in the owner-occupancy rate coefficient.

Column (4) from Table 4 and Column (4) from Table 5 offer a sharp comparison. After addressing the endogeneity concern using fixed effects, owner-occupancy rate is a strong driver of mayoral election turnout but not presidential election turnout. The same sharp contrast also holds using the log of election turnout as the dependent variable. After adding in census block fixed effects, the homeownership effect on presidential election turnout vanishes in Column (7) and (8) in Table 5. These results confirm the hypothesis that, compared to renters, homeowners are more likely to vote in local elections due to economic incentives.

Then I pool mayoral and presidential elections data together to statistically test the coefficient difference between $\hat{\beta}_m$ and $\hat{\beta}_p$. I define a mayoral election dummy – equaling 1 for data from mayoral elections and 0 for data from presidential elections. I regress voter turnout on the mayoral dummy, the owner-occupancy rate, the interaction term between the two, other controls, fixed effects, and interactions between the mayoral dummy and each of the aforementioned controls. Table 6 summarizes the pooled regression results. As before, Columns (1) - (3) use turnout rate as the dependent variable and Columns (4) - (6) use the log of turnout rate as the dependent variable. The first row reports the coefficient on the interaction term between the mayoral dummy and the owner-occupancy rate. It measures the difference between $\hat{\beta}_m$ and $\hat{\beta}_p$. A positive estimate indicates that $\hat{\beta}_m$ is larger than $\hat{\beta}_p$. In all specifications, the interaction coefficient estimates remain positive and significant. This implies owner-occupancy rate is a strong driver of mayoral election turnout as opposed to presidential election turnout. The second row reports the coefficient on the mayoral dummy. Unsurprisingly, the coefficient is negative in all columns since mayoral elections have lower turnout rates. The third row presents the effect of owner-occupancy rate on presidential turnout. The coefficient remains statistically indistinguishable from zero in most columns. It becomes negative in Columns (2) and (5), but the effect goes away with a full set of controls in Columns (3) and (6). Results from Table 6 confirm the previous findings.

Last, I provide further evidence of using the fixed effects model to obtain identification. As discussed in the introduction, one interpretation of controlling for block-level unobserved attributes is that it indirectly absorbs personal confounders that contribute to residential sorting. If individuals sort into narrowly defined geographic areas, a coarsely defined geographic fixed effect should not be able to correct the endogeneity bias. This implies that, by substituting for geographically refined fixed effects with geographically broader fixed effects, the endogenous correlation between owner-occupancy rate and voter turnout should emerge. In this paper, the most refined geographic level is a census block. The mean size of census blocks in the data is around 1,000 residents, approximately the size of a U.S. city neighborhood (roughly two to three street blocks in densely developed urban areas). A natural choice for a geographic unit that is coarser than a census block yet more refined than a city is the census tract. A census tract is usually composed of three to four census blocks. At the tract level, uneven distribution of public goods and amenities also exists. Massey (2001) demonstrates that the tract-level "Dissimilarity index"[18] is relatively high in both Philadelphia and Chicago (Seattle is not covered in their study). The census tract fixed effects are therefore well suited for an alternative specification using a broader geographic measure. Beyond census tracts, I use city fixed effects as the coarsest geographic unit control.

Table 7 summarizes the findings with different levels of fixed effects. Panel A uses only the mayoral election data while Panel B uses only the presidential election data. The full set of block-level controls is included in all specifications. Only the coefficient estimates on owner-occupancy rate are reported. Starting from Panel A, Column (1) presents the raw correlation between owner-occupancy rate and mayoral election turnout rate conditional on block attributes. Adding election cycle fixed effects in Column (2) brings limited change. Column (3) further adds city fixed effects and the coefficient largely remains stable. Column (4) substitutes city fixed effects with census tract fixed effects, and the coefficient declines almost by nearly half, from 0.117 to 0.060. Column (5) substitutes census tract fixed effects with the more refined census block fixed effects, and the coefficient declines further. The same pattern also holds from Columns (6) -

[18] "Dissimilarity index" measures is the relative number of Blacks who would have to change geographic units so that an even Black-White spatial distribution could be achieved

(10) using the log turnout rate as the dependent variable. In Panel B, which uses presidential elections data, we also observe a similar declining pattern in the coefficients. Most importantly, as I control for the most refined census block fixed effects, the correlation between owner-occupancy rate and presidential election turnout vanishes. This pattern is in line with the assumption that refined geographic fixed effects can absorb unobserved personal confounders associated with residential sorting. The residential sorting effect is weaker at a broader geographic level, thereby the broader fixed effects fail to capture the endogenous relationship between homeownership and voter turnout.

4.2. Control Function Approach

The CF specification is given by Equation (6). Compared to the fixed effects model, the CF approach allows the block-level confounders to be time-varying. As discussed in Section 2.2, in this more flexible specification owner-occupancy rate should have limited impact on presidential election turnout. By assuming $\beta_p = 0$, I can identify β_m which measures the causal impact of owner-occupancy rate on mayoral election turnout. To increase estimation precision, I pool the mayoral election records from all cities and election cycles together. City by election cycle fixed effects are included into the models to account for heterogeneity across city and time.

Table 8 presents the CF approach estimation results. Column (1) only includes owner-occupancy rate and neighborhood turnover rate as controls. The signs on both coefficients are consistent with previous findings from the fixed effects model. Column (2) adds presidential election turnout rate to control for unobserved confounders. Consistent with my expectation, the inclusion of presidential election turnout drives down the coefficient estimate on owner-occupancy rate. The coefficient estimate on neighborhood turnover rate also decreases. Similarly, the coefficient on the neighborhood turnover rate measures the differential impact of mobility on voter turnout in mayoral elections versus in presidential elections (scaled by $\frac{\gamma_m}{\gamma_p}$). How to interpret this coefficient depends on the assumptions one is willing to make. My goal of including the neighborhood turnover rate is to control for mobility, therefore I do not make extra assumptions to interpret this coefficient. The coefficient estimate on presidential election turnout is close to 0.45.

Column (3) adds the remaining census block controls into the regression. The coefficients for these controls are not reported for the same reason as with the neighborhood turnover rate. As the result shows, conditional on mobility and other controls, the coefficient estimate on owner-occupancy rate $\hat{\beta}_m$ is 0.05. This is close to the 0.046 estimate obtained from the fixed effects model (Table 4 Column (4)). The coefficient estimate on presidential election turnout remains stable. Similar patterns are found in the non-linear models in Columns (4) - (6) that use log of mayoral election turnout as dependent variable. Results from the CF approach confirm the previous findings from the fixed effects model.

4.3. Robustness Check

I conduct four robustness checks for both the fixed effects model and the CF approach. First, I drop census blocks located near major universities (U Penn, U Chicago and U of Washington) because those blocks contain a higher proportion of international students (non-U.S. citizens) which cannot be distinguished from the non-internal student population based on the census block level data available. I also drop blocks located in non-residential areas such as harbors, factories or parks because demographic estimates from those areas may be less accurate. Those are just a handful of blocks and their elimination does not change any result. Second, to eliminate any concern that the results are driven by outliers or measurement error, I limit the sample to blocks with turnout rate between 5% and 95% in all elections. Third, I run the specifications using the original full sample including census blocks with turnout larger than one. Fourth, I include controls for each block's racial composition in the full model to alleviate concerns that the homeownership effect may be caused by systematic differences in turnout across racial groups. None of the coefficients on racial composition controls are significant, and all other results hold.

For the fixed effects model I conduct one more robustness check. The identification assumption in the fixed effects model is that the block-level confounders are time-invariant. One may argue that gentrification may invalidate the time-invariant assumption. Therefore, I drop census blocks defined as been gentrified between 2000 and 2016 (about 10% of the original sample) by the *Governing* website[19]. All point estimates remain the same.[20]

5. Conclusion

This paper documents the fact that homeowners are more likely to vote in U.S. mayoral elections compared to renters living in the same community. There are two potential channels leading to this result. First, homeowners are more financially invested in the community by way of owning a house. Their financial stake in a community makes them more eager to participate in local political processes to promote policies that protect the property value. Renters, on the other hand, do not have similar stakes in a community and face fewer economic incentives to vote.

[19] The *Governing* website classifies gentrification at Census track level. A Census tract is considered to be gentrified if (a) the tract's median household income and median home value fall within the bottom 40th percentile of all tracts within a metro area in 2000; (b) the tract increases into the top third percentile for both inflation-adjusted median home values and percentage of adults with bachelors' degrees in 2016. Details please refer to http://www.governing.com/gov-data.

[20] I present all the robustness check tables in the appdendix of the online version of this paper. Link: http://boqianjiang.weebly.com/uploads/1/1/2/9/112968739/boqian_voting_paper.pdf

Second, homeownership also creates frictions on household mobility. Therefore, homeowners receive longer utility flow from local public goods consumption. Controlling for the mobility channel, my results suggest that the financial incentive alone is the driver of homeowners' tendency to vote in mayoral elections. Identification in this paper is obtained by (a) using a fixed effects model to control for unobserved confounders; (b) directly modeling the confounders using a control function approach. Both approaches use presidential election turnout outcome as the counterfactual. Since national-level elections and polity initiatives are less relevant to property value and local public good provision, the difference in tendency to vote between homeowners and renters should diminish in presidential elections.

The homevoter hypothesis literature shows home-voters and lease-voters vote differently in local political processes due to housing price capitalization effects (Fischel, 2001; Dehring et al., 2008; Ahlfeldt and Maennig, 2015). The findings in this paper further beg the question of whether uneven turnout between homeowners and renters in U.S. local elections causes biased policies favoring homeowners (e.g., using more strict zoning laws to keep housing value appreciation). To gain support, local political candidates may design policies to please the high turnout group. This may result in policies that likely fail to maximize total social welfare. Kahn (2011) provides some evidence that homeowner cities are more likely to block new housing development based on data from California. Future work could marshal evidence as to whether uneven turnout between homeowners and renters affects the design of local policy or policy outcomes.

References

Gabriel M. Ahlfeldt. Blessing or curse? Appreciation, amenities and resistance to urban renewal. *Regional Science and Urban Economics*, 41(1), January 2011: 32–45.

Gabriel M. Ahlfeldt and Wolfgang Maennig. Homevoters vs. leasevoters: A spatial analysis of airport effects. *Journal of Urban Economics*, 87(C), 2015: 85–99.

Andr´e Blais. *To vote or not to vote?: The merits and limits of rational choice theory*. University of Pittsburgh Pre, 2000.

Eric Brunner and Jon Sonstelie. Homeowners, property values, and the political economy of the school voucher. *Journal of Urban Economics*, 54(2), September 2003: 239–257.

Eric Brunner, Jon Sonstelie, and Mark Thayer. Capitalization and the Voucher: An Analysis of Precinct Returns from California's Proposition 174. *Journal of Urban Economics*, 50(3), November 2001: 517–536.

Barry C. Burden. The dynamic effects of education on voter turnout. *Electoral Studies*, 28(4), 2009. Special issue on The American Voter Revisited: 540 – 549.

Kerwin Kofi Charles and Jr. Stephens, Melvin. Employment, wages, and voter turnout. *American Economic Journal: Applied Economics*, 5(4), 2013: 111–43.

Carolyn A. Dehring, Craig A. Depken II, and Michael R. Ward. A direct test of the homevoter hypothesis. *Journal of Urban Economics*, 64(1), July 2008: 155–170.

Denise DiPasquale and Edward L. Glaeser. Incentives and Social Capital: Are Homeowners Better Citizens? *Journal of Urban Economics*, 45(2), March 1999: 354–384.

Gary V. Engelhardt, Michael D. Eriksen, William G. Gale, and Gregory B. Mills. What are the social benefits of homeownership? Experimental evidence for lowincome households. *Journal of Urban Economics*, 67(3), May 2010: 249–258.

William A Fischel. *The homevoter hypothesis*. Harvard University Press, 2001.

Edward L. Glaeser, Joseph Gyourko, and Raven E. Saks. Why Have Housing Prices Gone Up? *American Economic Review*, 95(2), May 2005: 329–333.

Zoltan Hajnal and Jessica Trounstine. Where turnout matters: The consequences of uneven turnout in city politics. *The Journal of Politics*, 67, 5 2005: 515–535.

Zoltan Hajnal and Jessica Trounstine. America's uneven democracy: race, turnout, and representation in city politics, 2010.

Christian A.L. Hilber and Christopher Mayer. Why do households without children support local public schools? Linking house price capitalization to school spending. *Journal of Urban Economics*, 65(1):74–90,

Thomas M Holbrook and Aaron C Weinschenk. Campaigns, mobilization, and turnout in mayoral elections. *Political Research Quarterly*, 2013.

Chang-Tai Hsieh and Enrico Moretti. Why do cities matter? local growth and aggregate growth. Working Paper 21154, National Bureau of Economic Research, May 2015.

Yannis M Ioannides and Kamhon Kan. Structural estimation of residential mobility and housing tenure choice. *Journal of Regional Science*, 36(3), 1996: 335–363.

Matthew E. Kahn. Do liberal cities limit new housing development? evidence from California. *Journal of Urban Economics*, 69(2), 2011: 223 – 228.

Alexander Keyssar. *The right to vote: The contested history of democracy in the United States*. Basic Books, 2009.

Vardges Levonyan. What led to the ban on same-sex marriage in california?: Structural estimation of voting data on proposition 8. Technical report, Mimeo, Harvard University, 2013.

Douglas S Massey. Residential segregation and neighborhood conditions in us metropolitan areas. *America becoming: Racial trends and their consequences*, 1(1), 2001: 391–434.

Scott Minkoff. Political engagement in a public goods context: The effect of neighborhood conditions and schools on local election turnout. *Available at SSRN 2416914*, 2014.

Raven Molloy, Christopher L. Smith, and Abigail Wozniak. Internal Migration in the United States. *Journal of Economic Perspectives*, 25(3), Summer 2011: 173–96.

J Eric Oliver, Shang E Ha, and Zachary Callen. *Local elections and the politics of small scale democracy*. Princeton University Press, 2012.

Franois Ortalo-Magne and Andrea Prat. On the Political Economy of Urban Growth: Homeownership versus Affordability. *American Economic Journal: Microeconomics*, 6(1), February 2014: 154–81.

Sherwin Rosen. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), Jan.-Feb. 1974: 34–55.

Stuart S Rosenthal. A residence time model of housing markets. *Journal of Public Economics*, 36(1), 1988: 87–109.

Stephen Ross and John Yinger. Sorting and voting: A review of the literature on urban public finance. In P. C. Cheshire and E. S. Mills, editors, *Handbook of Regional and Urban Economics*, volume 3 of *Handbook of Regional and Urban Economics*, chapter 47. Elsevier, 1999: 2001–2060.

Annika Smits and Clara H. Mulder. Family dynamics and first-time homeownership. *Housing Studies*, 23(6), 2008: 917–933.

Rachel Milstein Sondheimer and Donald P. Green. Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science*, 54(1), 2010: 174– 189.

Charles M. Tiebout. A Pure Theory of Local Expenditures. *Journal of Political Economy*, 64, 1956: 416.

Raymond E Wolfinger and Steven J Rosenstone. *Who votes?*, volume 22. Yale University Press, 1980.

Jeffrey M Wooldridge. Control function methods in applied econometrics. *Journal of Human Resources*, 50(2), 2015: 420–445.

John Yinger. Hedonic markets and sorting equilibria: Bid-function envelopes for public services and neighborhood amenities. *Journal of Urban Economics*, 86(C), 2015: 9–25.

Appendix

A. Data Sources:

A.1. Election data:

Philadelphia election data is retrieved from Philadelphia City Commissioners website: http://www.philadelphiavotes.com/en/resources-a-data/ballot-box-app.

Chicago election data is retrieved from Chicago Board of Election Commissioners website: http://www.chicagoelections.com/en/election3.asp.

Seattle election data is retrieved from King County's county website: http://www.kingcounty.gov/depts/elections/elections/past-elections.aspx.

In Chicago, the precinct definition changed between year 2011 and 2012. To append three rounds Chicago election data together I manually build a correspondent between old and new Chicago precinct using GIS software. The methodology is the same as the one I use to build correspondence between 2000 census block groups and 2010 census block groups. Detail of the GIS work is laid out below.

A.2. Census data:

Most of the census block groups level demographic data is not available on the American Fact Finder (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). I retrieve the data from ACS Summary file on Census server and clean them following the manual book provided by Census. Detail about the Summary file and retrieving guidance is available at http://www.census.gov/programs-surveys/acs/data/tools/summary-file-retrieval-tool.html.

The Voting Age Population by Citizenship and Race (CVAP) data is a separate product by Census. It’s available at https://www.census.gov/rdo/data/voting_age_population_by_citizenship_and_race_cvap.html.

B. Census blocks correspondence file building procedure:

For each of the three cities, I use GIS software to build the correspondence between 2000 census blocks (census block groups) and 2010 Census blocks in following steps:

(1) Obtain the 2000 and 2010 census block groups shapefiles from Census website https://www.census.gov/geo/maps-data/data/tiger-cart-boundary.html

(2) Overlap the 2000 and 2010 shapefiles together in the GIS software.

(3) For each 2000 census block groups, the GIS software calculates its overlapping with all 2010 census block groups. Then based on the calculation the GIS software will assign a number to indicate what percentage of a 2000 census block groups belongs to a 2010 census block groups. This generates a correspondence file for the given city.

Table 1: Mean and dispersion measure of election turnout rate from three U.S. cities

City	Election cycle	N	Mayoral election turnout		Presidential election turnout	
			μ	σ / μ	μ	σ / μ
Philadelphia	1	1129	0.42	0.26	0.62	0.19
	2	1129	0.25	0.36	0.66	0.24
	3	1129	0.16	0.43	0.60	0.26
Chicago	1	1874	0.24	0.45	0.53	0.32
	2	1874	0.26	0.42	0.61	0.26
	3	1874	0.33	0.33	0.55	0.32
Seattle	1	377	0.40	0.35	0.66	0.27
	2	377	0.45	0.31	0.70	0.23
	3	377	0.41	0.34	0.69	0.26

Note: The election turnout rates are measured at census block level. This election panel covers mayoral and presidential elections held between 2003 and 2013 for Philadelphia, Chicago and Seattle. The elections are grouped into three election cycles and the election cycles are four years apart. Each election cycle contains a mayoral election and a presidential election. The detail of the election grouping is summarized in Table 2.

Table 2: Election panel grouping list and merge with Census data

Election cycle	Mayoral election	Presidential election	Census Data
1	2003 Philly, Chicago 2005 Seattle	2004 All Three Cities	Average of 2000 Decennial Census and ACS 2005-2009 5-year-estimates
2	2007 Philly, Chicago 2009 Seattle	2008 All Three Cities	ACS 2006-2010 5-year-estimates
3	2011 Philly, Chicago 2013 Seattle	2012 All Three Cities	ACS 2010-2014 5-year-estimates

Note: The election panel covers mayoral and presidential elections held between 2003 and 2013 for Philadelphia (Philly), Chicago and Seattle. Data merge between election data and Census data is based on 2010 Census blocks definition.

Table 3: Summary statistics for election turnout and owner–occupancy rate

City	Election cycle	N	Mayoral election turnout mean / median (sdv)	Presidential election turnout mean / median (sdv)	Owner-occupancy rate mean / median (sdv)
Philadelphia	1	1129	0.42 / 0.43 (0.11)	0.62 / 0.63 (0.12)	0.59 / 0.61 (0.21)
	2	1129	0.25 / 0.25 (0.09)	0.66 / 0.65 (0.16)	0.57 / 0.58 (0.23)
	3	1129	0.16 / 0.16 (0.07)	0.60 / 0.59 (0.16)	0.54 / 0.55 (0.23)
Chicago	1	1874	0.24 / 0.23 (0.11)	0.53 / 0.53 (0.17)	0.48 / 0.45 (0.23)
	2	1874	0.26 / 0.34 (0.11)	0.61 / 0.61 (0.16)	0.49 / 0.47 (0.24)
	3	1874	0.33 / 0.31 (0.12)	0.55 / 0.53 (0.18)	0.47 / 0.44 (0.24)
Seattle	1	377	0.40 / 0.41 (0.14)	0.66 / 0.69 (0.18)	0.49 / 0.50 (0.25)
	2	377	0.45 / 0.46 (0.14)	0.70 / 0.71 (0.17)	0.49 / 0.50 (0.26)
	3	377	0.41 / 0.43 (0.14)	0.69 / 0.72 (0.18)	0.49 / 0.50 (0.25)

Note: The unit of observation is census blocks. The election panel covers mayoral and presidential elections held between 2003 and 2013 for Philadelphia (Philly), Chicago and Seattle. The owner-occupancy rate data is obtained from the Census.

Table 4: Fixed effects model estimation results using mayoral election panel data

Dependent variable	turnout rate				log (turnout rate)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Owner-occupancy rate	0.197^{***}	0.158^{***}	0.078^{***}	0.047^{***}	0.705^{***}	0.516^{***}	0.372^{***}	0.229^{***}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.04)	(0.06)	(0.07)
% Moved into current residence within past 12 month		-0.167^{***}	-0.048^{**}	-0.012		-0.812^{***}	-0.127	0.023
		(0.01)	(0.02)	(0.02)		(0.08)	(0.08)	(0.08)
Household median income (Million $)				-0.767^{***}				-3.085^{***}
				(0.12)				(0.50)
% High school and some college among pop>25				-0.088^{***}				-0.288^{***}
				(0.02)				(0.10)
% College above among pop>25				0.083^{***}				0.403^{***}
				(0.02)				(0.11)
% Married-household				0.020				0.121^{*}
				(0.02)				(0.07)
% Adults 30 to 60				0.209^{***}				0.833^{***}
				(0.02)				(0.10)
% Adults 60 above				0.301^{***}				1.251^{***}
				(0.03)				(0.11)
R-squared	0.24	0.26	0.06	0.10	0.17	0.19	0.05	0.09
Election cycle fixed effects	3	3	3	3	3	3	3	3
City fixed effects	3	3	–	–	3	3	–	–
Census block fixed effects	–	–	3369	3369	–	–	3369	3369
Observations	10107	10107	10107	10107	10107	10107	10107	10107

Note: The mayoral election panel data is from Philadelphia, Chicago and Seattle between 2003 and 2013. The unit of observation is census blocks. Robust standard errors clustered at census blocks in parenthesis. *** Significant at 1%, ** significant at 5%, * significant at 10%.

Table 5: Fixed effects model estimation results using presidential election panel data

Dependent variable	turnout rate				log (turnout rate)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Owner-occupancy rate	0.096***	0.070***	-0.016	-0.026	0.175***	0.116***	-0.010	-0.022
	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)
% Moved into current residence within past 12 month		-0.108***	-0.035	0.008		-0.251***	-0.057	0.018
		(0.02)	(0.02)	(0.02)		(0.07)	(0.05)	(0.05)
Household median income (Million $)				-0.857***				-1.515***
				(0.15)				(0.30)
% High school and some college among pop>25				-0.025				-0.097
				(0.03)				(0.07)
% College above among pop>25				0.120***				0.162**
				(0.03)				(0.07)
% Married-household				-0.065***				-0.122***
				(0.02)				(0.05)
% Adults 30 to 60				0.261***				0.430***
				(0.03)				(0.07)
% Adults 60 above				0.345***				0.625***
				(0.03)				(0.07)
R-squared	0.10	0.10	0.07	0.11	0.06	0.07	0.05	0.08
Election cycle fixed effects	3	3	3	3	3	3	3	3
City fixed effects	3	3	–	–	3	3	–	–
Census block fixed effects	–	–	3369	3369	–	–	3369	3369
Observations	10107	10107	10107	10107	10107	10107	10107	10107

Note: The presidential election panel data is from Philadelphia, Chicago and Seattle between 2003 and 2013. The unit of observation is census blocks. Robust standard errors clustered at census blocks in parenthesis. *** Significant at 1%, ** significant at 5%, * significant at 10%.

Table 6: Fixed effects model results using pooled mayoral and presidential election panel data

Dependent variable	turnout rate			log (turnout rate)		
	(1)	(2)	(3)	(6)	(7)	(8)
Owner-occupancy rate * Mayoral dummy	0.078	0.078	0.073	0.446	0.446	0.250
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.06)
Mayoral election dummy	-0.296	-0.296	-0.018	-0.883	-0.883	-0.996
	(0.00)	(0.00)	(0.03)	(0.01)	(0.01)	(0.10)
Owner-occupancy rate	-0.000	-0.029	-0.026	-0.024	-0.119	-0.022
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
% Moved into current residence within past 12 month		-0.002	0.008		0.020	0.018
		(0.02)	(0.03)		(0.06)	(0.06)
Household median income (Million $)		-0.812	-0.857		-2.300	-1.515
		(0.12)	(0.16)		(0.35)	(0.33)
% High school and some college among pop>25		-0.057	-0.025		-0.192	-0.097
		(0.02)	(0.03)		(0.07)	(0.07)
% College above among pop>25		0.101	0.120		0.282	0.162
		(0.02)	(0.03)		(0.08)	(0.08)
% Married-household		-0.023	-0.065		-0.000	-0.122
		(0.02)	(0.02)		(0.05)	(0.05)
% Adults 30 to 60		0.235	0.261		0.631	0.430
		(0.02)	(0.03)		(0.07)	(0.08)
% Adults 60 above		0.323	0.345		0.938	0.625
		(0.03)	(0.04)		(0.08)	(0.07)
R-squared	0.70	0.71	0.77	0.62	0.63	0.71
Election cycle fixed effects	3	3	3	3	3	3
Census block fixed effects	3369	3369	3369	3369	3369	3369
Mayoral dummy interaction with election cycle fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Mayoral dummy interaction with other slope variables	–	–	Yes	–	–	Yes
Mayoral dummy interaction with block fixed effect	–	–	Yes	–	–	Yes
Observations	20214	20214	20214	20214	20214	20214

Note: The election panel data is from Philadelphia, Chicago and Seattle between 2003 and 2013. The unit of observation is census blocks. Robust standard errors clustered at census blocks in parenthesis. *** Significant at 1%, ** significant at 5%, * significant at 10%.

Table 7: OLS estimation result comparison across different fixed effects models

Dependent variable	turnout rate					log (turnout rate)				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Mayoral Election										
Owner-occupancy rate	0.093***	0.091***	0.117***	0.060***	0.047***	0.219***	0.209***	0.356***	0.244***	0.229***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.04)	(0.04)	(0.04)	(0.07)
R-squared	0.22	0.25	0.33	0.51	0.10	0.18	0.20	0.26	0.53	0.09
Panel B: Presidential Election										
Owner-occupancy rate	0.151***	0.145***	0.123***	0.082***	-0.026	0.284***	0.271***	0.213***	0.172***	-0.022
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
R-squared	0.17	0.19	0.21	0.45	0.11	0.11	0.12	0.14	0.50	0.08
Other slope variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Election cycle fixed effects	–	3	3	3	3	–	3	3	3	3
City fixed effects	–	–	3	–	–	–	–	3	–	–
Census track fixed effects	–	–	–	1177	–	–	–	–	1177	–
Census block fixed effects	–	–	–	–	3369	–	–	–	–	3369
Observations	10107	10107	10107	10107	10107	10107	10107	10107	10107	10107

Note: The election panel data is from Philadelphia, Chicago and Seattle between 2003 and 2013. The unit of observation is census blocks. Panel A uses election panel data from mayoral elections. Panel B uses election panel data from presidential elections. Robust standard errors clustered at census blocks in parenthesis. *** Significant at 1%, ** significant at 5%, * significant at 10%.

Table 8: Control function estimation results using election panel data

Dependent variable	Mayoral election turnout			log (Mayoral election turnout)		
	(1)	(2)	(3)	(6)	(7)	(8)
Owner-occupancy rate	0.151***	0.120***	0.051***	0.490***	0.391***	0.131***
	(0.01)	(0.00)	(0.01)	(0.04)	(0.02)	(0.03)
% Moved into current residence within past 12 month	-0.161***	-0.112***	0.044***	-0.787***	-0.563***	-0.197***
	(0.01)	(0.01)	(0.01)	(0.08)	(0.05)	(0.04)
Presidential election turnout		0.453***	0.450***			
		(0.12)	(0.16)			
Log (Presidential election turnout)					0.890***	0.882***
					(0.06)	(0.07)
R-squared	0.49	0.76	0.79	0.39	0.75	0.78
City by election cycle fixed effects	9	9	9	9	9	9
Other slope controls	No	No	No	No	No	No
Observations	10107	10107	10107	10107	10107	10107

Note: The election panel data is from Philadelphia, Chicago and Seattle between 2003 and 2013. The unit of observation is census blocks. Robust standard errors clustered at census blocks in parenthesis. *** Significant at 1%, ** significant at 5%, * significant at 10%.

Chapter 2.

Separating Selection From Spillover Effects: Using the Mode to Estimate the Return to City Size

Hugo Jales
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, New York, 13244-1020
hbjales@syr.edu

Boqian Jiang
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, New York, 13244-1020
bjiang03@syr.edu

Stuart S. Rosenthal
Maxwell Advisory Board Professor of Economics
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, New York, 13244-1020
ssrosent@maxwell.syr.edu

Abstract

We develop a new method to identify and control for selection when estimating the productivity effects of city size. For single peaked factor return distributions, selecting out low-performing agents has no effect on modal productivity but reduces the CDF evaluated at the mode. Spillovers from agglomeration have the reverse effect. This holds regardless of whether selection arises from the decision to participate or location choice. Estimates based on law firm productivity, wages for married women and wages for full-time men all confirm that selection contributes to urban productivity and that doubling city size causes productivity to increase by 1-2.5 percent.

JEL Codes: R00 (General Urban, Rural, and Real Estate Economics)

Key Words: Agglomeration economies, selection, city size, mode versus mean

1. Introduction

A challenge for all studies that seek to estimate the productivity effects of agglomeration and city size is the need to separate out selection from spillover effects (see Rosenthal and Strange (2004) and Combes and Gobillon (2015) for reviews). This arises because cities are expensive places in which to live, work and operate a business (e.g. Rosenthal and Strange (2012), Combes et al. (2012), Black et al (2014)) so that only the most productive workers and companies participate – a threshold effect. It also arises because cities may attract unusually talented individuals who thrive on the intensity of urban life – a migration effect (e.g. Glaeser and Mare (2001), Rosenthal and Strange (2008), Combes et al. (2008), de La Roca (2017)). Both forms of selection contribute to higher levels of productivity in cities, confounding efforts to identify the causal impact of agglomeration on individual productivity.[1] Building off recent work by Combes et al. (2012), this paper develops a new method that identifies the presence and nature of selection while yielding estimates of the causal effect of city size on productivity.

Combes et al. (2012) argue that the presence or absence of selection effects can be identified by examining the shape of the observed factor return distribution. They note that if companies drop out when factor productivity is below a common threshold, selection left-truncates the observed distribution of returns. Assuming further that productivity thresholds increase with city size, they examine whether truncation is more prevalent among larger cities, using data on manufacturing plants in France. They fail to find evidence of such patterns and conclude from this that higher manufacturing productivity in larger French cities arises primarily from spillover effects and not from selection.

[1] Common approaches to deal with the endogenous selection of workers and companies into different sized cities include the use of pseudo-random experiments (e.g. Ahlfeldt et al (2015)) and instrumental variables (e.g. Rosenthal and Strange (2008)). Nevertheless, the confounding effects of selection remain challenging and pseudo-experiments and instrumental variable approaches often offer solutions that do not extend beyond the immediate study.

This paper extends the Combes et al. (2012) model in ways that yield a more general approach to controlling for selection effects. Our model applies to settings in which the latent factor return distribution is single peaked with a well-defined mode. This is characteristic of wage and earnings patterns, for example, where it is common to model the underlying distribution of returns as log normal. In such instances, if selection disproportionately culls out lower performing units to the left of the mode, then the CDF evaluated at the mode will be reduced. For sufficiently single-peaked factor return distributions, however, the modal level of productivity is highly robust to selection whereas selecting out lower performing agents pushes up the observed mean return.[2] Implementing these ideas points to two complementary regressions. The first regresses the CDF of the modal factor return on log city size while the second regresses log modal factor return on log city size. Evidence of a negative city size effect in the first regression indicates that selection occurs disproportionately to the left of the mode. In such instances, the second regression should also yield an estimate of the return to city size that is below that of the mean return. Moreover, with sufficiently precise estimates of the mode (in a sense to be clarified later), the second regression yields estimates of the return to city size that are robust to selection effects.

The model above can be used to evaluate the presence, nature and impact of selection both when selection arises from threshold effects and when selection arises from migration (sorting). However, the two sources of selection entail different modeling assumptions that in some instances affect interpretation and robustness. In the threshold model, we focus on the

[2] Selecting out low-performing agents will also push up the median return. However, throughout the paper we emphasize comparisons between modal and mean values rather than the median. Partly this is because the first regression described below does not extend to median values since, by definition, the CDF evaluated at the median is always 50 percent, and also because the vast majority of studies in the agglomeration literature have focused on mean returns to city size.

decision to participate, as when a firm survives and remains in business or when an individual chooses to work, while treating location choice as exogenous. In the migration model, we focus on city choice and the possibility that talented individuals may sort into larger cities; in this model we treat the decision to participate as exogenous. Both sources of selection yield the same qualitative predictions as described above. However, the threshold model is a closed city model in which selection contributes to higher productivity in larger metropolitan areas because of rising operating costs or other mechanisms (e.g. competition as in Combes et al (2012)) that push up the productivity threshold. The migration model, in contrast, is an open city framework. In this model, sorting of skilled individuals into larger cities equivalently implies that lower-skilled individuals disproportionately sort into smaller cities. Nevertheless, while the interpretation associated with the two forms of selection (threshold versus migration) is different, the anticipated patterns from the two-part regression framework above are the same provided the factor return distribution is sufficiently single-peaked.[3]

Relative to our approach, it is worth noting that Combes et al (2012) treat selection as arising from a common threshold that increases with city size. Our threshold model, in contrast, allows for heterogeneity of threshold levels across individuals and establishments within a given city, effectively treating the threshold as a random variable. The presence of such heterogeneity will make it more difficult to discern evidence of truncation in the factor return distribution. Our approach based on the CDF of the mode provides a simple but revealing way to identify the presence and nature of selection that allows for within-city heterogeneity. Moreover, in the special case where all threshold-based selection is to the left of the mode – even allowing for

[3] We show later in the paper that the first regression patterns based on the CDF evaluated at the mode are especially robust to the source of selection (threshold versus migration). The second regression based on modal return is also robust provided the tendency for individuals to migrate into larger cities increases monotonically with skill but not at too high a rate relative to the degree to which the factor return distribution is single-peaked.

within-city threshold heterogeneity – the rate at which the CDF at the mode declines with city size is an exact measure of the extent of selection. In the more general case where selection occurs throughout the factor return distribution, a decline in the CDF at the mode indicates that selection occurs disproportionately below the mode and the rate at which that increases with city size provides a lower bound estimate of the extent of selection. In each of these cases, the modal value of productivity or wage will be highly robust to selection effects provided the factor return distribution is sufficiently single-peaked.

For the migration model, we show later in the paper that interpretation of our first regression depends on two key modeling assumptions. The first as above is that the latent aggregate factor return distribution is single-peaked with a well-defined mode. The second is that the tendency for individuals to select into larger cities increases monotonically with skill. Provided these conditions are met, migration of skilled individuals into larger cities causes the CDF evaluated at the modal factor return to decrease with city size. Moreover, and again as above, provided the return distribution is sufficiently single-peaked relative to the rate at which skilled individuals select into larger cities, modal productivity will be robust to selection in contrast to the mean.

This paper is the first we are aware of to use modal productivity when measuring the return to city size. The vast literature on agglomeration economies has instead focused almost exclusively on mean returns (see Rosenthal and Strange (2004) and Combes and Gobillon (2015) for reviews). Whether our estimates based on the modal return are impactful depends in part on the degree to which the mode is of intrinsic interest for the outcome measure being considered. In instances where factor return distributions are symmetric and single-peaked, this is straightforward as the mode, mean and median are all alike. Where factor return distributions are single-peaked but skewed, the mode is also often as informative a summary measure of the central tendency of the return distribution as the median and mean, but for all three measures context matters. On the other hand, using the mode to identify the presence and nature of selection is robust to any context for which the single peak condition is plausible.[4]

We use three data sets to illustrate our approach and to provide new estimates of the nature of selection and return to city size. To highlight threshold effects, we use law firm productivity for all law firms across the United States, drawing on establishment level data from Dun and Bradstreet. A key modeling assumption for this example is that entry and exist costs for law firms are low but annual operating costs are high. Moreover, we assume that lawyers initially have imperfect knowledge of their own ability and discover through experience whether they can profitably run their own law firm or are better off working for an established company. Law firms that are not sufficiently productive eventually close as annual loses persist. As formalized later in the paper, under these conditions, threshold effects should be more pronounced among older law firms since only the most productive companies survive. Assuming further that operating costs are higher in larger cities, we expect this pattern to increase with city size. These sharp priors provide an opportunity to check whether the first regression based on CDF evaluated at the mode is successful at identifying the presence and nature of selection.

[4] Our emphasis on using the mode as a measure of the central tendency of a distribution in the presence of selection effects has antecedents in earlier work by Lee (1989). Lee showed that under certain conditions, the mode from a truncated distribution is a consistent estimate of the conditional mean from the original, non-truncated distribution. As with Combes et al (2012), Lee (1989) focused on the case where the point of truncation is known and common across agents. Our work is also broadly related to the modal regression literature in statistics that construct statistical models by exploiting different properties of the mode (Huang et al. 2013; Yao and Li, 2014; Chen et al., 2016). That literature, however, does not consider the robust nature of the mode in the presence of selection. In economics, focus of modal values is rare. Cardoso and Portugal (2005) show that modal wage is a better measure of the central tendency of the underlying wage distribution when there is collective bargaining. Bound and Krueger (1991) and Hu and Schennach (2008) discuss how to use mode to account for certain forms of reporting errors when measuring the distribution income. Our approach is also related to the "identification at infinity" models in Chamberlain (1986), Lewbel (2007), and D'Haultfoeuille and Maurel (2013). These models assume that selection effects shrink to zero as certain key control variables approach "infinity". In these models, however, selection is based on one or more control variables whereas selection in our paper is based on the dependent (outcome) variable.

We also test our model using individual wage rates for married full-time working white women, drawing on data from the 5 percent file of the 2000 U.S. census (obtained from IPUMS). It is well-established that married female labor supply is highly elastic so the decision to work full time is relevant to threshold effects (e.g. Heim (2007); Blau and Kahn (2007)). [5] Contributing to this view, Black et al. (2014) argue that higher commuting costs in large cities discourage married women from working. Married female labor supply decisions may also affect choice of metropolitan area as in Costa and Kahn (2000). For these reasons, wage patterns for married women are likely to be driven by a combination of threshold and migration effects. In the analysis to follow, our models based on female wage rates are estimated stratifying women into skilled (college degree or more) and low-skilled (high school degree or less) individuals. That is because labor supply elasticity likely differs for high- and low-skilled married women and for that reason selection effects associated with city size may differ as well, although the direction of any such differences is less clear. In this context, our model has potential to reveal whether selection effects related to city size are more pronounced for skilled versus low-skilled married women.

To illustrate migration effects, we focus on prime age, 25 to 54 year old, full-time white male workers in the U.S. Consistent with extensive work in the labor literature, for this sample, we treat the decision to work as inelastic and exogenous (e.g. Heim (2007); Blau and Kahn (2007)). Under the assumption that labor supply is exogenous, selection effects for this group are likely driven disproportionately by migration and location decisions. Data for this exercise were also taken from the 2000 Census.

[5] Based on 1999-2001 CPS data, Blau and Kahn (2007) find that the elasticity of annual working hours with respect to *own* log wage is 0.357 for the married women and 0.046 for the married men. Moreover, the elasticity of annual working hours with respect to *spouse's* log wage is -0.192 for the married women and -0.006 for the married men.

For all three exercises, estimates based on the CDF evaluated at the mode yield robust and compelling evidence that selection contributes to productivity in larger cities. For law firms, the modal return from doubling city size is approximately 1 percent as compared to 1.7 percent at the mean. For this application, however, we also find that the estimated return to city size at the mode is sensitive to the bandwidth with which the mode is measured. As discussed more fully later in the paper, this highlights a power issue and limitation of our approach: in addition to requiring that the underlying latent distribution be single-peaked, sample size must be large enough to yield a sufficiently precise estimate of the mode. For the wage applications, the observed distributions are especially single-peaked and the sample sizes are large. In these applications, for both married women and prime age men, the sample modes are robust to reasonable alternative choice of bandwidth as are the estimated returns to city size at the modal values. For skilled women, doubling city size increases modal wage by 2.3 percent compared to 4.3 percent at the mean; for low-skilled women, selection effects are largely absent and doubling city size increases wage by roughly 3 percent at both the mode and the mean. For men, doubling city size increases the modal wage by roughly 2.5 percent compared to 4.5 percent at the mean. As with married women, evidence of selection effects is largely absent for low-skilled men.

We proceed as follows. The next section outlines our selection model, first for heterogeneous threshold effects and then for migration. Section 3 describes the data and present summary statistics. Section 4 discusses how to measure the mode. Sections 5 presents the results and Section 6 concludes.

2. Model

This section presents our modeling framework. We begin with the influence of agglomeration economies on the distribution of worker productivity in large versus small cities in the absence of selection effects. The model is then extended to allow for threshold-based selection and selection arising from migration.

2.1. Productivity spillovers from city size

Suppose initially that there are no selection effects that influence the distribution of productivity in large versus small cities. Instead, the only force that causes productivity distributions to differ across metropolitan areas are spillovers arising from city size. To simplify, we assume two different size cities, denoted as 0 for small cities and 1 for large cities. Productivity spillovers from agglomeration increase productivity in larger cities.

Let individual worker productivity be denoted by y, and let $f_0(y)$ and $f_1(y)$ represent the distribution of productivity among individuals in small and large cities, respectively. Cities are assumed to be large enough that $f_0(y)$ and $f_1(y)$ are approximately continuous on y, and worker productivity in a size-0 city depends only on a worker's intrinsic level of skill. If agglomeration economies increase productivity by a common factor for all workers, $f_1(y)$ shifts to the right relative to $f_0(y)$. If instead, the returns to city size increase with skill, possibly because more talented workers are better able to take advantage of large city opportunities, then this would create a "dilation effect" (Combes et al., 2012) causing $f_1(y)$ to become right skewed with an elongated right tail. Allowing for both effects, for a given individual, productivity in a larger city is given by,

$$y_1 = \beta_0 + \beta_1 y_0 \tag{2.1}$$

In expression (2.1), β_0 measures the common productivity boost for all workers in a larger city while $\beta_1 > 1$ would imply that the returns to city size increase with worker skill. As in Combes et. al (2012), expression (2.1) specifies spillover effects in a linear form for which shift and dilation effects preserve an individual's productivity rank within a given city. Under these conditions, the cumulative distribution function (CDF) for productivity up to a given skill level, y_0, is the same in each city, F_0 and F_1,

$$F_0(y_0) = F_1(y_1(y_0)) \tag{2.2}$$

Substituting for y_1 from expression (2.1) and taking derivatives, the relationship between large and small city productivity densities is given by,

$$f_1(y) = \frac{1}{\beta_1} f_0\left(\frac{y-\beta_0}{\beta_1}\right) \tag{2.3}$$

Our most important modeling assumption in the empirical work to follow, referred to as Assumption 1, is given by:

Assumption 1: *$f_0(y)$ is single peaked with a well defined mode at an interior location.*

In conjunction with spillover effects as modeled in (2.3), this assumption has important implications for the shape of productivity density functions in large versus small cities. To illustrate, we took 10,000 random draws of log(y_0) from a normal distribution (with standard deviation of 0.4), mirroring assumptions in the labor and agglomeration literatures that typically treat wage and earnings distributions as log-normal. For illustrative purposes, we also set $\beta_0 = 0.5$ and $\beta_1 = 1.3$. Figure 1 then traces out the resulting productivity density functions for large cities (the dashed red line) and small cities (the solid black line).

Notice that for large cities, the density function is right shifted with an elongated right tail (right skewed) relative to the density function for small cities. The large city density is also flatter, with a lower density for any given level of productivity, and a right shifted mode. A positive value for β_0 shifts the large city distribution along the x-axis by β_0 units while preserving its shape. This is apparent from (2.1) and (2.3). In (2.1), the derivative of y_1 with respect to β_0 is 1 while from (2.3) the large city density with $\beta_1 = 1$ is $f_0(y - \beta_0)$. Observe, however, that the mode in Figure 1 shifts from 0.85 in small cities to 1.55 in large cities even though β_0 is just 0.5. The additional rightward shift in the mode is a consequence of the dilation effect arising from $\beta_1 > 1$ which draws the mode further to the right, although not immediately apparent from a casual viewing of (2.3). As is evident in the figure, the mode in the large city density is also not as pronounced relative to a smaller city. This also arises from $\beta_1 > 1$, which flattens the density function by shifting mass from the center of the distribution into the increasingly elongated right tail, and bearing in mind that the density function must always integrate to 1.

2.2. Threshold effects

Consider now the influence of threshold effects that contribute to selection and which differ across agents within a given city. We assume that the latent productivity distributions are identical in small and large cities but threshold effects are more pronounced in larger metropolitan areas. For simplicity, small city residents are described below as participating in the labor market with probabiltiy 1 regardless of skill, or $\pi_0(y) = 1$, where $\pi_0(y)$ is the probability of participating. If in the large city $\pi_1(y)$ is also constant with $\pi_1(y) = p < 1$, then the selection process is random and $f_1(y) = f_0(y)$. More relevant for our context, however, is the possibility that in large cities participation increases with skill, which we formalize as our second core modeling assumption:

Assumption 2: *In large cities, the probability of participating in the labor market increases monotonically with skill,* $\partial\pi_1(y)/\partial y > 0$.

Assumption 2 captures the tendency for operating costs tend to be higher in larger cities and/or the environment more competitive (as in Combes et al (2012)). For that reason, weaker companies are more likely to drop out in larger cities relative to the experience in smaller metropolitan areas. Anlaogously, because commuting costs tend to be higher in larger cities, Assumption 2 captures the sense that lower productivity workers are more likely to drop out of the labor force in larger cities relative to smaller metropolitan areas (see Black et al (2014) for related discussion).

Allowing for heterogeneous threshold effects as above, expression (2.3) becomes,

$$f_1(y) = \frac{\pi_1(y)}{\beta_1 c} f_0\left(\frac{y-\beta_0}{\beta_1}\right) \tag{2.4}$$

where $c = \int \pi_1(u) f_0(u) du$. In (2.4), note that $\pi_1(y) < 1$ reduces the density for a given level of y in the larger city, while c scales the density up by the inverse share of agents that participate (firms or individuals), ensuring that the density function integrates to 1.

We illustrate the qualitative effects of threshold-based selection in Figures 2 and 3 using the same simulated data as for Figure 1, first without and then with spillovers. In Figure 2, we set $\beta_0 = 0$ and $\beta_1 = 1$, consistent with the absence of agglomeration economies. The $\pi_1(y)$ function is specified such that $\pi_1(y)$ increases up to a value of 1 at the mode of the latent distribution (at y = 0.85) and remains at 1 thereafter.[6] Imposing these features, ten percent of the simulated work

[6] More precisely, we set $\pi(y) = - 0.27 + 1.5y$ for $y \leq 0.85$ and $\pi(y) = 1$ for $y \geq 0.85$. Specified in this manner,

force is selected out of the large city labor market, all of whom have skill levels to the left of the mode. The important point to recognize in Figure 2 is that even though all selection occurs to the left of the mode, selection steepens the slope of the large city density function on both sides of the mode while also increasing the height of the mode. Together, these effects cause the modal level of productivity in the density function to become more pronounced.

Figure 3 illustrates the combined influence of threshold and spillover effects. In this instance we set β_0 and β_1 to the values used in Figure 1 and specify $\pi_l(y)$ as in Figure 2. In Panel A, notice that the influence of threshold effects is difficult to discern relative to the pattern in Figure 2. That is because dilation associated with $\beta_1 > 1$ flattens and right-skews the distribution causing the mode to become less pronounced. This offsets the tendency for threshold effects to accentuate the mode. On the other hand, because in this example all selection is to the left of the mode in the large city population, the CDF evaluated at the mode must be reduced relative to the CDF at the mode in the small city distribution. This is readily apparent in Panel B which shows that the corresponding CDFs evaluated at the respective small and large city modes are 0.34 and 0.27.

The patterns in Figures 2 and 3 are based on an extreme selection process for which all selection is to the left of the mode. Nevertheless, the patterns highlight two principles that apply in the more general settings in which selection occurs throughout the productivity distribution.

Proposition 1: *Given Assumptions 1 and 2, if selection occurs disproportionately to the left of the mode in the large city distribution, then the CDF evaluated at the mode declines with city size while the reverse is true if selection occurs disproportionately to the right of the mode.*

$\pi_l(y) = 0$ for the lowest level of y in the simulated sample and approaches 1 asymptotically from below at $y = 0.85$.

This proposition motivates our first regression described in the Introduction and points to a simple, robust way to identify whether selection occurs more to the left or right of the mode of a latent productivity distribution. Moreover, in the special case where selection occurs only to the left (or right) of the mode, the difference in CDF evaluated at the mode for large versus small cities is an exact measure of the extent of selection.

The second principle implicit in Figures 2 and 3 is that if (i) the latent distribution is sufficiently single peaked with a well defined mode and (ii) the selection process that governs the manner in which π_1(y) changes with y is not too extreme in nature, threshold-based selection will not affect the value of the mode in the observed productivity distribution. This points to our second proposition and related regression.

Proposition 2: *Provided that the mode in the underlying latent single-peaked distribution is sufficiently well defined and the selection process is not too extreme in nature, selection will have a small effect on the modal value in the observed productivity distribution. Under these conditions, the observed difference in modal values between large and small cities approximately measures the productivity spillover effect from city size.*

In considering this proposition, it is important to emphasize that extreme forms of selection would shift the mode in an observed productivity distribution. If, for example, $\pi_1(y) = 0$ for $y < y^*$, where y^* is above the mode of the latent distribution, then selection would increase the mode in larger cities. If instead, however, $\pi_1(y)$ increases gradually and monotonically with y, it is straightforward from Figure 2 to show that selection would not affect the mode provided the latent distribution is sufficiently single peaked.

A third principle implicit in Figures 2 and 3 is that provided $\pi_1(y)$ increases monotonically with y, selection will typically have less impact on the mode of the observed distribution than the mean or median. This brings us to our third proposition.

> **Proposition 3:** *If the underlying latent distribution is sufficiently single-peaked with a well defined mode (in a manner to be clarified), and if $\pi_1(y)$ increases monotonically with y, selection will typically have less impact on the mode of the observed distribution than the mean or median.*

There are many contexts in which one would expect $\pi_1(y)$ to increase monotonically with y and for which the underlying latent distribution would be expected to be single-peaked. In such instances, Proposition 3 emphasizes that the mode is typically less sensitive to selection and yields a more robust measure of the underlying relationship than the mean or median.

2.3. Migration effects

Consider next the influence of migration as the source of selection effects. In this instance, we assume a common aggregate single-peaked (latent) productivity distribution from which individual workers sort into two types of cities, small (size 0) and large (size 1). In this setting, $\pi(y)$ represents the probability that a worker with skill level y chooses to locate in the larger city. As with the threshold model, if $\pi(y)$ equals a constant p, the selection process is random and $f_1(y) = f_0(y)$. In this instance, selection would not affect the CDF evaluated at the modes in small and large cities. A more realistic scenario, however, is that $\pi(y)$ increases in a smooth, monotonic fashion with y, analogous to Assumption 2 above, and consistent with the view that higher skilled individuals are more likely to select into larger cities. This would also simultaneously reduce skill levels in smaller urban areas. Nevertheless, all three propositions outlined above still hold.

To clarify, consider first an extreme but illustrative selection process. We set $\pi(y) = 0$ for $y \leq y^*$ and $\pi(y) = 1$ for $y > y^*$, where y^* is an interior point in the aggregate distribution. Specified in this manner, all workers below y^* sort into the small city while all of those above y^* sort into the large city. Figures 4a and 4b highlight implications of these conditions using the same simulated data as above. The key difference between the figures is whether y^* is below or above the modal level of skill in the aggregate distribution, denoted by y_m and equal to 0.85 as before.

In Figure 4a we set y^* equal to 0.65 so that $y^* < y_m$. This causes the small city density function (in the top portion of Panel A) to increase monotonically with y with a mode equal to $y^* = 0.65$. The large city density, in contrast (in the top portion of Panel B), declines monotonically from a modal value equal to $y_m = 0.85$. In the lower portions of each panel, notice also that the CDF evaluated at the mode in the small city equals 1 since all workers have productivity below y^*, while the CDF for the large city must be less than 1 since the mode is at an interior location. In Figure 4b we instead set y^* equal to 1.0 so that $y^* > y_m$. This causes the small city mode to equal y_m while the large city mode becomes y^*. Under these conditions, the CDF evaluated at the large city mode collapses to 0 and the corresponding CDF for the small city is positive but less than 1. The important point to emphasize from these patterns is that regardless of whether y^* is above or below y_m, the CDF evaluated at the mode declines with city size. Provided our core modeling assumptions 1 and 2 hold, therefore, Proposition 1 is robust to threshold and migration effects as alternate sources of selection.

Consider now a more realistic characterization of migration for which $\pi(y)$ increases with y in a smooth, gentle and monotonic fashion. To illustrate the influence of such a process, in Figures 5 and 6 we again display large and small city productivity density functions using the same simulated data as before. In both figures, we also specify $\pi(y)$ so that the likelihood of locating in a large city increases linearly with y at rate $0.1y$ and with $\pi(y)$ set equal to 0.5 for the least skilled individual in the sample. This also ensures that $\pi(y) = 1$ for the most skilled individual in the sample.[7] In Figure 5a, spillover effects are set to zero with $\beta_0 = 0$ and $\beta_1 = 1$ in expression (2.4), while in Figure 5b we allow for spillover effects using the same specification as for Figure 1.

Focusing first on Figure 5a, it is evident that the specified migration process has little effect on modal productivity values, similar to the pattern in Figure 2 for threshold effects. Migration does, however, have noteworthy effects in Figure 5a. Relative to large cities, migration increases the height of the density function evaluated at the small city mode and steepens the slope of the density function on either side of the small city mode. This is opposite from the influence of threshold effects in Figure 2, and reinforces the principle that the height of the density function evaluated at the mode and the slopes of the density function on either side of the mode are not necessarily reliable indicators of selection effects even when the core modeling assumptions 1 and 2 hold. This conclusion is made even stronger when the influence of productivity spillovers is taken into account. In the upper panel of Figure 5b, dilation arising from $\beta_1 > 1$ again flattens and right-skews the productivity density function in large cities relative to small cities, further masking the influence of migration (as in Figure 3). In the lower panel of Figure 5b, however, which plots the CDFs for the small and large city productivity

[7] Specified in this manner, 60 percent of workers in the simulated sample sort into the large city.

distributions, the respective CDFs evaluated at the modes are 0.38 and 0.31. Once again, the CDF evaluated at the mode declines with city size, consistent with Proposition 1.

Returning to the upper panel of Figure 5b, observe also that the modal productivity values for small and large cities are 0.85 and 1.55, respectively. Because the underlying latent distribution is single peaked and the selection process is not too extreme, the difference in modal productivity between large and small cities is largely unaffected by selection and reflects primarily the effect of city size on productivity. This reinforces Proposition 2. More generally, because migration shifts mass to the right in the large city productivity density function relative to the small city, that will tend to increase the spread between large and small city means (and medians). This once again suggests that the mode is less sensitive to selection relative to the mean and median of the underlying productivity density functions.

2.4. How sensitive is the mode to selection?

The results above require that the underlying aggregate (latent) density function for the outcome measure is sufficiently single peaked, and that the selection process is not too extreme. This section formalizes when these conditions are met.

Suppose that selection effects are present but agglomeration economies are not. Then $\beta_0 = 0$, $\beta_1 = 1$, and the conditional density in (2.4) becomes,

$$f_1(y) = \frac{\pi_1(y)}{c} f_0(y) \quad . \tag{2.5}$$

The question we seek to answer is by how much selection may shift the mode of the conditional density $f_1(y)$ relative to the unconditional density $f_0(y)$. Since $f\,(y)$ is assumed to be twice differentiable and single peaked, its slope at the mode is zero. Differentiating (2.5) with respect to y and setting the derivative to zero, the modal value for y (denoted by y_m) in the conditional density $f_1(y)$ must satisfy,

$$\frac{\pi'(y)}{\pi(y)} = -\frac{f_0'(y)}{f_0(y)} \tag{2.6}$$

Expression (2.6) indicates that at the mode, a small change in y yields equal magnitude but opposite signed percentage changes in the selection probability and the density of y. Multiplying both sides of (2.6) by y this can be expressed as an elasticity condition,

$$\xi_{\pi,y} = -\,\xi_{f_0,y} \tag{2.7}$$

where $\xi_{\pi,y} \approx \frac{\%\Delta\pi(y)}{\%\Delta y}$ and $\xi_{f_0,y} \approx \frac{\%\Delta f_0(y)}{\%\Delta y}$.[8] Expression (2.7) says that at the modal value of the conditional density function, the elasticity of the selection probabilty is equal to minus the elasticity of the latent density. If the selection probability increases (or decreases) monotonically with y, along with the asumed shape of the latent distribution, expression (2.7) will be satisfied at a single value for y, ensuring that the conditional density is also single peaked. This is assured because both the density and selection functions are assumed to be log concave.[9]

Figure 6 illustrates these principles. The upper panel displays a twice differentiable single peaked density function and a linear monotonically increasing selection function with a vertical intercept at the origin. The lower panel plots the corresponding values for $-\,\xi_{f_0,y}$ and $\xi_{\pi,y}$. In the case where y is normally distributed, it is straightforward to show that $-\,\xi_{f_0,y} = y(y - y_m)/\sigma^2$ with a slope of $(2y - \mu)/\sigma^2$ that increases at a rate of $2/\ \sigma^2$. In this instance, $-\,\xi_{f_0,y}$

[8] The elasticities above express the percent change along the vertical axis in response to a percent change along the horizontal axis. This is the inverse of familiar demand and supply elasticities. Nevertheless, the elasticities in (2.6) are specified as above because y is the exogenous determinant of f and π.

[9] This result follows from arguments in Saumard and Wellner (2014) and An (1996). We assume that $f_0(y)$ and $\pi(y)$ are both log-concave functions with $f_1(y)$ as their product scaled by a normalizing constant, c. From Proposition 3.2 in Saumard and Wellner (2014), the product of two log-concave functions is log-concave so that $f_1(y)$ is also log-concave. Note also that Proposition 2 in An (1996) indicates that a random variable y is distributed in a log-concave fashion if and only if its density function is strongly unimodal. Together, these principles imply that $f_1(y)$ is unimodal and there is a unique value for y (in its feasible range) for which (2.7) holds.

initially declines from zero at the origin to a minimum at $y = y_m/2$, and increases monotonically thereafter, taking on a value of 0 at the mode and positive values thereafter. As drawn in the upper panel, the selection function has a constant unit elasticity up to the point where $\pi(y) = 1$, after which $\xi_{\pi,y} = 0$. The elasticity plots in the lower panel must therefore intersect to the right of y_m, indicating that selection shifts the mode of the conditional density function to the right. If instead, the selection function was flat, then $\xi_{\pi,y} = 0$ for all y and the elasticity plots intersect at y_m. In this instance, selection is random and does not affect the mode. Alternatively, if selection declines monotonically with y, expression (2.6) still holds but the mode in the conditional density function will shift to the left.

Two final comments remain when considering the viability of using the mode to test and control for selection effects. First, sample size must be large enough to yield sufficiently reliable estimates of the mode for purposes of evaluating the CDF at the mode and the impact of city size on modal productivity. This point is considered further in the empirical sections to follow. Second, the mode needs to be of intrinsic interest for the problem being considered. While these conditions will not always hold, they are met in many problems regularly considered in economics.

3. Data and Summary Statistics

3.1. Three datasets

This section describes the three datasets used to estimate the model above. In the first instance, we use old and newly established law firms to look for evidence of selection effects and to estimate the return to city size. As described in the Introduction, entry and exit costs are low for lawyers operating their own firms. Suppose also that lawyers only learn whether they can profitably operate their own firm from experience, and the returns from running a law firm are high if the venture is successful. Under these conditions, a wide range of lawyers may attempt to establish their own companies, including many who are less adept but do not realize their firms are likely to fail. This would reduce tendencies for threshold-based selection at the point of entry. After a few years, however, lawyers discover their type and weaker companies drop out so that threshold effects should be especially apparent among older companies. Moreover, with higher operating costs and a more competitive environment in larger cities, evidence of threshold-related selection and related differences between new and older law firms should increase with city size.

As also described in the Introduction, for full-time working married white women, it is plausible that both threshold and migration effects contribute to selection and higher observed wages in larger cities. Threshold effects, for example, could arise if longer commute times in larger cities discourage women from working (e.g. Black et al (2014), while migration effects could be associated with job market co-location challenges that draw skilled couples to larger cities (e.g. Costa and Kahn (2000)). In contrast, for full-time working white men, labor supply is highly inelastic. For this group, migration effects seem likely to be the dominant source of selection. The data used for each of these applications is described below.

3.2. Law firm establishment data from Dun & Bradstreet

We collect establishment-level data for all law firms in the United States (excluding Alaska and Hawaii) from the Dun & Bradstreet Million Dollar Database. The data provides information on establishment location, level of employment, sales, industry (SIC 8-digit code), year established, and other information. Compared to the Census data, an advantage of Dun &

Bradstreet database is that it provides a comprehensive coverage of small businesses including those with just one or two-workers.[10]

The data were collected in December 2016 and provide a snapshot of all law firms operating in the U.S. at that time. We use establishment-level sales per worker as a proxy for productivity. Based on the sales per worker measure, we trim the top and bottom 0.1% of the data to drop outliers.[11] Certain types of law offices may be more prevalent in large cities (e.g. corporate law). Because of concerns about selection stem from unobserved factors embedded in the error term, we pre-cleaned the data to difference out the average return for the primary classifications of law firms identified in the data.[12] This was done by regressing individual establishment sale per worker on dummy variables for each type of 8-digit law office recorded in by Dun and Bradstreet. We then added back to the residual the average sale per worker for general law offices/attorneys which account for 90% of the sample.

MSA size is measured using population estimates from the 2015 American Community Survey.[13] A key part of our empirical strategy is to measure the mode of the sales per worker distribution in each MSA. To make sure there are sufficient numbers of law firms present, we retain only MSAs for which all of the following conditions are satisfied: (i) more than 30 law firms age five or younger are present, (ii) more than 30 law firms over five years in age are present, and (iii) MSA total population is over 100,000. After cleaning the data as above, we are

[10] In our sample, there are 545,873 law establishments. Of these, 8.5% have one worker, 62.8% have two employees, 15.0% have three workers, and 13.5% have four or more workers. In comparison, in the 2012 Economic Census, there are 186,831 law establishments in the U.S. The main reason for the difference is that Census indicates that it does not "survey very small businesses". For details see the Census website: https://www.census.gov/programs-surveys/economic-census/about/faq.html.

[11] Similar trimming procedure is also used in Combes et al. (2008), Combes et al. (2012) and Gaubert (forthcoming).

[12] Based on SIC 8-digit codes, approximately 90% of the sample is coded as general law offices/attorneys. The remaining 10% of the sample is coded into more specialized classifications, including corporate law, family law, etc.

[13] Throughout the paper, the 2013 Office of Management and Budget metropolitan area delineations are used to define MSAs.

left with 239 MSAs. The total count of law firms in the sample is 545,873 firms. Of these, 74,079 firms are young, defined as five years or less in age, and 471,794 firms are old, defined as over five years in age.[14]

Table 1 Panel A presents summary statistics of sales per worker for all law firms sample and separately by age group (young and old) as well. Based on the 25th and 75th quantile, the majority of the sales per worker measures fall within the range of $60,000 and $85,000. Measured at the mean and different quantiles, old firms have higher sales per worker than the young firms, indicating that older law firms are more productive than younger companies.

Figure 7a provides kernel density plots of sales per worker for the all firms sample as well as for the different age groups. In each panel, the sales per worker distribution is singled peaked.[15] In Figure 7b, kernel density plots are provided again, stratifying each sample into small (population < 1m) and large (population > 2.5m) MSAs. The plots make clear that the distribution of sales per work in large cities is right-shifted as compared to small cities for in all three samples.

3.3. Married white female full-time workers in the 2000 Census

The sample of married female non-Hispanic white full-time workers (age 25-54) was obtained from the 2000 decennial census 5% public use micro sample (PUMS) from IPUMS

[14] Among the 545,873 establishments, age related information was missing in the D&B data for 67,358 establishments (12% of the sample). For roughly 200 of these firms, we searched the companies on the web by establishment name (which is also reported by D&B). In each instance, the establishments was over 5 years in age. For that reason, we classified all law firms in D&B with missing age information as over 5 years in age (i.e. as old establishments).

[15] There are also several spikes in the density estimation, indicating rounding errors in the sales per worker data. The rounding errors are likely to be caused by the fact that firms tend to report sales rounded by thousands of dollars. We will discuss how to deal with the rounding errors in Section 4.

(www.ipums.org)[16] Full-time workers were coded as those who report working at least 35 hours per week and 40 weeks per year.[17] Hourly wage was used as a proxy for productivity.

Individual hourly wage is calculated by dividing annual earnings by annual hours worked, where the later is given by weeks worked by usual hours worked per week. As above, we trim the top and bottom 1% of the sample based on hourly wages. Also analogous to above, we pre-clean the data by regressing individual wage on age fixed effects, education fixed effects, occupation fixed effects and industry fixed effects.[18] We retain the wage residual from each worker and calculate the adjusted wage for each worker by adding coefficient from the constant term to the wage residuals that restores the original sample mean. We clean the wage data for skilled (college degree or more) and low-skilled (high school degree or less) workers separately.[19]

MSA population size is estimated using the 2000 census.[20] We retain only those MSAs for which all of the following conditions are satisfied: (i) more than 100 married female non-Hispanic white workers with a college degree or more present, (ii) more than 100 married female non-Hispanic white workers with a high school degree or less present, and (iii) MSA total population is over 100,000. The data cleaning procedure leaves us a sample composed of

[16] We obtain the sample through the IPUMS website (Steven et al., 2015). Samples from Alaska and Hawaii are excluded. We only use non-Hispanic white workers sample because discrimination against certain ethnic groups may affect the observed wage distribution in a fashion that may be related to city size. We also restrict the sample to native-born.

[17] We focus on full-time workers in part to reduce measurement error when calculating hourly wages which is more pronounced among part-time workers. See Baum-Snow and Neal (2009) for related discussion.

[18] To be specific, there are 15 age fixed effects, 359 occupation fixed effects and 94 industry fixed effects. In the census, the most detailed version of occupation classification is at 6 digits, which is too refined that certain occupations do not have enough sample size to yield precise estimates of fixed-effects. Therefore, we choose to control for occupation fixed effects using 5-digit classification. As a robustness check, we find that controlling for occupation fixed effects at 4-digit or 6-digit level also yield similar results.

[19] This approach does not prevent the adjusted wage from being negative. And there are indeed a few instances that the adjusted wage is negative in our sample. It does not affect our analysis.

[20] The population estimate is obtained through the IPUMS website. Link: https://usa.ipums.org/usa-action/variables/MET2013#description_section

152,704 skilled married female workers and 153,168 low-skilled married female workers from 216 MSAs in the United States.

Table 1, Panel B provides summary statistics of adjusted hourly wage for the married female workers. Measured at the mean and each quantile, the adjusted hourly wage is higher among the skilled workers. Figure 8 Panel A and B present kernel density plots of the adjusted hourly wage for low- and high-skilled workers. The first thing to note is that both distributions are single-peaked. The density plot for the skilled workers (Panel A) also has a longer right tail and a larger variance as compared to the plot for low-skilled workers (Panel B). Splitting the samples into small and large MSAs, we reproduce the density plots in Panels C and D. For both groups of workers, the wage density plots for large cities is right-shifted and dilated as compared to the density plot for small cities.

3.4. Male full-time white workers in the 2000 Census

Male non-Hispanic white full-time workers (age 25-54) data is also drawn from the 5% PUMS of the 2000 decennial Census. These data are cleaned in the same way as for the married female workers. This leaves us with 383,728 workers with a college degree or more and 393,598 have a high school degree or less. These workers are spread 262 MSAs in the United States. Table 1, Panel C summarizes the adjusted hourly wage for the skilled (college degree or more) male workers and low-skilled male workers (high school degree or less).

Unsurprisingly, the skilled workers have higher adjusted hourly wage than the low-skilled group, both at various quantiles and also at the mean. Figure 9, Panels A and B display kernel density plots of the adjusted hourly wage for the two groups of male workers. In both panels, the aggregate adjusted wage distributions are single-peaked. Splitting the samples into small and large cities (Panels C and D, respectively), it is also evident that the wage density for large cities is also right-shifted and dilated as compared to the density plot for smaller cities, similar to the patterns for the married female sample.

4. Measuring the mode

Our estimation procedure requires that we measure the modal value of the outcome variables (e.g. sale/worker, wage) in each MSA. We illustrate how this is done using law establishment sale per worker data.

We first discretize the sales per worker distribution in each MSA by rounding the sales per worker values to the closest integral using a fixed bandwidth. The choice of rounding bandwidth will be discussed shortly. Then we define the modal sales per worker of each city as the rounded point that has the highest frequency in each MSA's discretized sales per worker distribution.

We use this method mainly for two reasons. First, discretizing sales per worker using a fixed bandwidth across cities ensures comparability across MSAs and different samples. Second, discretizing as above mitigates measurement error associated with rounding when the raw data are reported. As shown in Figure 7a, the density plot of sales per worker includes a number of spikes that likely arise from rounding in the reported values. Discretizing the data using a fixed bandwidth likely eliminates or at least greatly reduces rounding error associated with the reported values.[21]

[21] We exam the finite sample property of our approach using data simulation. We find that model estimates using our approach are consistent. Results available upon request. As another robustness check, we also identify the mode from kernel density estimation, allowing the bandwidth to vary across cities. We find that estimation results from the two methods are very similar.

A key part of the procedure above is the choice of bandwidth used to discretize the data. If the bandwidth is too narrow, the discretized distribution will converge towards a uniform distribution with a poorly defined mode. If the bandwidth is too wide, variation in discretized values will be so reduced that it will not be possible to discern meaningful patterns since all of the data would eventually be coded to a single cell. It is necessary, therefore, to select a bandwidth that balances these two extremes.

We begin by first documenting the inter-quartile range for the sales per worker distribution. In Table 1, Panel A, the inter-quartile range for the law firm sales per worker is roughly $20,000 to $25,000 for all three main samples, including all law firms, young and old. Next, we compare the density plots for sales per worker in small and large cities in Figure 7b. In all three panels in the figure (for all law firms, young and old), the difference in modal sales per worker between large and small cities is less than $20,000. This suggests that any bandwidth larger than $10,000 would likely not preserve enough variation in the data.

Figure 10a presents histograms of the aggregate sales per worker data using a $5,000 bandwidth. It is evident that there is a well-defined mode in the distributions for all three groups of firms (all law firms, young and old). Figure 10b, Panels B and C, provide analogous histograms using alternative bandwidths. When we decrease the rounding bandwidth to $2,500 in Panel B, many of the adjacent histograms have similar height and the mode is not longer well defined. When we increase the bandwidth to $7,500 in Panel C, the histograms become very thick and we lose considerable variation.

For the reasons above, we choose a $5,000 bandwidth to discretize the sales per worker data in each city.[22] Using that bandwidth, Table 2, Panel A summarizes modal sales per worker

[22] We discuss the robustness of the modal estimates to bandwidths from $4,000 to $6,000 later in the paper.

estimates across MSAs. The difference in the minimum and maximum modal sales per worker is about $20,000 for all three groups of firms, consistent with the plots in Figure 7b.

The same procedure as above was used to select the bandwidths when measuring the wage distributions for married women and prime age men. In both samples the bandwidth was set equal to $3.

For women, in Table 1, Panel B the interquartile range of adjusted wage is $10 for the skilled married female workers and $5 for the low-skilled female workers. From the distribution plots in Figure 8, Panels C and D, the difference in the modal adjusted hourly wage between the small and the large cities is less than $5.[23] Figure 11a provides histograms of the adjusted hourly wage based on a $3 bandwidth. It is evident that there is a well-defined mode for both low- and high-skilled workers. Distribution plots in Figure 11b based on bandwidths of $1 and $5 do not perform as well. With a $1 bandwidth the mode is not well defined while the $5 bandwidth eliminates much of the variation in the distribution. Table 2, Panel B summarizes the wage distribution using a $3 bandwidth. Across the 216 MSAs, the minimum and maximum modal wages are $15 and $30 for the skilled married women, and $12 and $18 for the low-skilled married women.

For prime age men, Table 2 Panel C summarizes modal adjusted hourly wage estimates based on the discretized data. The minimum and maximum modal adjusted wages are $21 and $39 for the skilled male workers and $9 to $21 for the low-skilled male workers. Figure 12a plots histograms of the adjusted hourly wages using a $3 bandwidth. All panels have a well-defined mode. As a comparison, Figure 12b provides histograms using $1 and $5 as the alternative bandwidths. Analogous to the patterns for women, the $1 bandwidth (Panel B) is too refined to

[23] We also discuss the robustness of the results by varying the bandwidth from $2 to $4 in the empirical section.

effectively define the mode and the $5 bandwidth (Panel C) is too wide to preserve variation in the wage measure.

5. Empirical Results

5.1. Old and young law firms: Threshold effects

Table 3 presents regression results based on the law firm sales per worker. All regressions are at MSA-level. Panel A reports the regression for all firms. Panel B reports the regression for young law firms and Panel C reports estimates for older firms. In all cases, column (1) reports estimates from the first regression that uses the CDF evaluated at the mode as the dependent variable. Column (2) reports estimates from the second regression that uses log sales per worker at the *mode* as the dependent variable. Column (3) reports estimates using log sales per worker at the *mean* as the dependent variable in and column (4) reports coefficient difference test between the modal return (column 2) and the mean return (column 3).

Recall from Proposition 1 in Section 2, that selecting out establishments disproportionally to the left of the mode will cause the CDF evaluated at the mode to decline in value. This pattern should be most evident in larger cities if selection effect increases with city size. Table 3 column (1) tests for these patterns by reporting results of OLS regressions of the CDF evaluated at the mode on log population of the MSA. Consider first Panel A which reports estimates for all firms (including both young and old), the coefficient on log population is -0.02 with a t-ratio of 3.81. This indicates that, for all firms together, doubling city size reduces the CDF evaluated at the mode by 2 percentage points. This estimate confirms that selection is present and that on net weaker firms are selected out in large cities.

As discussed in the Introduction, comparing old and newly established law firms can highlight the effect of threshold-based selection. In Panel B column (1), the coefficient on log population is positive and significant (0.0158 with a t-ratio of 2.81), indicating the absence of threshold-based selection among the young firms. In comparison, the coefficient estimate for old firms is -0.023 with a t-ratio of 2.73 in Panel C column (1). The sharp contrast between the young and old firm results demonstrates that threshold-based selection in large cities is concentrated among older companies. That is consistent with a view that lawyers may not initially have full information on their ability to operate a profitable law firm. Partly for that reason, we expect a broader distribution of productivity among young law firms as aspiring lawyer-entrepreneurs are tempted to run their own firm. Over time, weaker lawyers discover they are not profitable and exit, contributing to the sharper pattern of selection for older companies.

Table 3 column (2) and (3) report regression estimates of the elasticity of sales per worker with respect to city size. As discussed earlier, we expect estimates based on the mean to be upward biased when a selection effect is present. Estimates based on modal establishments should be largely free of threshold-based selection and lower for that reason. In Panel A, column (2) the elasticity of sale per worker with respect to MSA size is 1 percent when measured using the modal firms. The corresponding estimate at the mean (column 3) is 1.69 percent. From the coefficient difference test in column (4), the return to city size measured at the mean is 0.6 percentage points higher than the return to city size measured at the mode. This pattern is consistent with our prior that selection effects should upward bias estimates of the return to city size when measuring based on the mean relative to the mode. The coefficients for older companies in Panel C are quite close numerically to the all firms sample in Panel A. For younger companies, however, a very different pattern is present. In Panel B, the coefficient based on the modal firms in column (1) is 3.05 percent, notably higher than corresponding values of 1.74 based on the mean (column 3). These patterns are in sharp contrast to the patterns for older establishments.

Estimates in Table 3 are based on discretized distributions using a \$5,000 bandwidth. To examine the robustness of the results with respect to bandwidth choice, we also estimated our models several additional times varying the bandwidth from \$4,000 to \$6,000 in \$200 increments. Figures 12a and 12b plot the resulting estimates.

The three panels in Figure 13a (for all law firms, young and old) report coefficient plots for the first step regression with the CDF evaluated at the mode as the dependent variable. From all three panels, the coefficient estimates largely remain stable when we vary the rounding bandwidths from \$4,000 to \$5,500. Within that range, the coefficient estimates for the all firms and old firms sample are less than zero (denoted by the red dashed horizontal line), while the coefficient estimates for the young firms are consistently larger than 0. Based on the 95% confidence interval, the differences are also statistically significant. Beyond the \$5,500 range, coefficient estimates become unstable in all three panels. The general pattern from this exercise is that the results based on CDF evaluated at the mode is largely stable with respect to rounding bandwidth choice.

Figure 13b present analogous coefficient plots for the elasticity of *modal* return to city size. The red dashed line in each panel represents the elasticity of *mean* return to city size from the corresponding sample. In this instance, it is apparent that the elasticity estimate of the modal return is sensitive to bandwidth choice. One explanation for this pattern is that the true elasticity of modal return with respect to city size is likely no larger than 5 percent given previous estimates in the literature. Relative to that value, over or underestimating the mode by even a few percentage points would be substantial relative to the anticipated return from a doubling of city size. This highlights a power issue for the law firm second regression. In contrast, the first step regression based on the CDF evaluated at the mode is much less sensitive to noise in the modal estimates as is apparent from the results discussed above.

Summarizing, the different estimates in Table 3 support the anticipated view that threshold-based selection will tend to cause weaker companies to drop out over time. Controlling for threshold effects, estimates suggest that doubling city size increases modal law firm productivity by roughly 1 percent. This is on the lower side of many estimates reported in the literature, where recent reviews suggest that most estimates are between 2 and 5 percent (Rosenthal and Strange (2004); Combes and Gobillon (2015)). However, it is also worth noting that nearly all estimates to date have focused on manufacturing and we are not aware of any that have been based exclusively on law firms.

5.2. Married full-time working women: Threshold and migration effects

Table 4 reports results based on the married female wage data. Panel A displays estimates based on the skilled (college degree or more) female workers sample while Panel B displays results for low-skilled workers.

Column (1) in each panel reports estimates based on the CDF evaluated at the mode. In Panel A, the coefficient on log MSA population is -0.02 with a t-ratio of 3.02. This suggests that, among the skilled married women, selection effects in larger cities disproportionally drive less productive individuals out of the full-time labor market. The presence of selection implies that the return to city size measured at the mean should be upward biased and higher than the return measured at the mode. These predictions are confirmed in Panel A, columns (3) and (4). The wage elasticity with respect to city size is 2.3 percent measured at the mode and 4.3 percent measured at the mean. Both estimates are statistically different from zero and from each other. This later point is confirmed in column (4) which reports a formal test of the difference between the two estimates. These patterns are consistent with our prior that failing to account for selection upward biases estimates of the return to city size.[24]

Consider next the results in Table 4, Panel B for low-skilled (high school degree or less) married women. In column (1), the estimated effect of log population on the CDF evaluated at the mode is -0.007 and not statistically different from zero. This suggests that selection is largely absent which further suggests that the return to city size measured at the mean and the mode should be similar. This prediction is confirmed in Panel B, columns (2) and (3). The elasticity of return to city size is 3.8 percent measured at the *mode* and 3.9 percent measured at the *mean*. Both estimates are statistically different from zero while column (4) confirms that the two estimates are not statistically different from each other.

Results in Table 3 were obtained for a discretized wage distribution using a $3 bandwidth. Figure 14a plots coefficient estimates based on the CDF evaluated at the mode for bandwidths ranging from $2 to $4 in $0.20 increments. In Panel A, the coefficient estimate on log population is smaller than zero (the horizontal red dashed line) and remains stable throughout different bandwidth. In Panel B, for the low-skilled population, the pattern is also consistent with the results in Table 3; in this case, the coefficient estimates are not statistically different from

[24] Recall from Figure 6, Panel C, that the kernel density plot for married female wages in large cities is dilated relative to the density plot for workers in small cities. Given the discussion in Section 2.1, this pattern implies that productivity spillovers in large cities may be disproportionally beneficial to the more productive individuals. Therefore, it is worth to emphasize that the elasticity estimate from Panel A column (2) is the productivity return to city size for the modal workers. Since mode measures the most frequent value in a distribution, the modal adjusted hourly wage is a meaningful central tendency measure in our study. Therefore, we believe that the 2.3 percent estimated modal return is an important and credible measure of spillover effects as compared to the mean, especially given the presence of selection effects.

zero in most instances.[25] Figure 14b present coefficient plots for the second step wage elasticity at mode. Different from the law firm sale per worker sample, in this instance the coefficient estimates are stable across the different bandwidths and close to estimates in column 2 of Table 3.

Summarizing, for college educated married women, there is compelling evidence that selection effects in larger cities disproportionally drive less productive workers out of the full-time market. Consistent with that pattern, doubling city size increases wages for college educated women by 2.3 percent measured at the *mode* and 4.3 percent measured at the *mean*. For the low-skilled married women evidence of selection effects associated with city size are largely absent, and doubling city size increases wage by roughly 3.8% at both the *mode* and the *mean*.

5.3. Male full-time workers: Migration effects

As discussed earlier, because male labor supply is very inelastic, migration is likely the dominant mechanism by which selection affects wage distributions. Table 5 reports estimates for this sample.

Focus first on Panel A which reports estimates for college educated men. In column (1), the impact of log population on the CDF evaluated at the mode is -0.008 with a t-ratio of 1.79. This is consistent with migration of more productive workers into larger cities. Consistent with that pattern, the wage elasticity with respect to city size is 2.5% when measured at the *mode* (in column 2) and 4.5% when measured at the *mean* (in column 3). In column 4, that difference in

[25] Although there is a visible drop in the coefficient estimate using around the bandwidth of $3.5, it is likely to be caused by limited variation in the modal wage estimates. Recall from Table 1 Panel B, the majority of the adjusted hourly wage for the low-skilled workers fall within the range of $12 to $18. Therefore, any rounding bandwidth that is large than $3 may yield modal estimates that have limited variation across cities.

estimates is significant, providing further evidence that among college educated men, unusually productive individuals tend to sort into larger cities.

Results for low-skilled men are presented in Panel B of Table 5. Findings mirror those for married women. In column (1), notice that there is no evidence that the CDF evaluated at the mode changes with respect to city size, consistent with an absence of sorting by unobserved skill into larger cities. In the absence of systematic sorting, the return to city size measured at the mean and at the mode should be similar. In Panel B, columns (2) and (3), the elasticity estimate is 4.4% measured at the *mode* and 3.6% measured at the *mean*. Although the point estimate of the modal return is higher than at the mean, the difference is statistically significant at only the 10% level (in column 4). In addition, the two estimates are not systematically different from each other for plausible alternative bandwidths used to discretize the data as shown in Figure 14.

Figure 15a presents coefficient estimate plots based on the CDF regression using different bandwidth. It is evident that the coefficient estimates in both panels remain stable and consistent with the results in Table 5 column (1). Figure 15b presents coefficient estimate plots for the elasticity of return to city size measured at the mode. A very stable pattern also emerges. Especially, Panel B shows that the modal return to city is not statistically different from the mean return to city size for the low-skilled workers.

To summarize, the patterns for male full-time workers indicate that among college educated individuals, migration tends to draw unusually productive individuals to larger cities. The elasticity of wage with respective city size is 2.5% measured at the *mode* compared to 4.5 percent measured at the *mean*. For low-skilled male full-time workers, there is no evidence of migration-related selection.

6. Conclusion

This paper develops a new method to identify and control for selection when estimating the productivity effects of city size. Different from previous papers, we emphasize that for single peaked factor return distributions, selecting out low-performing agents has no effect on modal productivity but reduces the CDF evaluated at the mode. Spillovers from agglomeration have the reverse effect. We show that these patterns hold regardless of whether selection arises from the decision to participate or location choice.

We estimate our model using three different data sets, each of which highlights different features of the approach. This includes establishment-level data for newly formed and older law firms, wages for full-time working married women, and wages for full-time working (prime age) men. Results from all three exercises yield robust and compelling evidence that selection contributes to urban productivity. The exception is for women and men with high school or less education. For that group, evidence of selection effects is largely absent.

Our results confirm that in many instances, failing to control for selection leads to overestimates of the returns to city size. For prime age, college educated men, for example, doubling city size increases productivity by 4.5 percent evaluated at the mean but just 2.5 percent evaluated at the mode.

Our approach can be applied to other contexts provided the underlying modeling assumptions are met. Those key features include that the underlying latent productivity distribution is sufficiently single-peaked with a well-defined mode, tht the selection process is not too extreme, and that the observed mode can be estimated with sufficient precision. Our method is also especially salient when modal values of the outcome measure are of intrinsic interest. However, even when that is not the case, the CDF evaluated at the mode can be used to test for the presence and nature of selection.

References

Ahlfeldt, G., S.J. Redding, D.M. Sturm, and N. Wolf (2015). The Economics of Density: Evidence from the Berlin Wall. *Econometrica* 83(6), 2127-2189.

An, Mark Yuying (1996). Log-Concave Probability Distributions: Theory and Statistical Testing. *Duke University Dept of Economics Working Paper* No. 95-03: 9-10

Behrens, K., Duranton, G., & Robert-Nicoud, F. (2014). Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy*, 122(3), 507-553.

Behrens, K., and F. Robert-Nicoud (2015). Agglomeration Theory with Heterogeneous Agents. G. Duranton, J. V. Henderson and W. Strange (eds), *Handbook in Regional and Urban Economics*, Amsterdam (Holland), Elsevier Press.

Black, D. A., Kolesnikova, N., & Taylor, L. J. (2014). Why do so few women work in New York (and so many in Minneapolis)? Labor supply of married women across US cities. *Journal of Urban Economics*, 79, 59-71.

Blau, F. D., & Kahn, L. M. (2007). Changes in the labor supply behavior of married women: 1980–2000. *Journal of Labor Economics*, 25(3), 393-438.

Baum-Snow, N., & Pavan, R. (2013). Inequality and city size. *Review of Economics and Statistics*, 95(5), 1535-1548.

Baum-Snow, N., & Neal, D. (2009). Mismeasurement of usual hours worked in the census and ACS. *Economics Letters*, 102(1), 39-41.

Bosquet, Clement and Pierre-Philippe Combes (2017). Sorting and agglomeration economies in French economics departments. *Journal of Urban Economics*, 101, 27-44.

Cardoso, A. R., & Portugal, P. (2005). Contractual wages and the wage cushion under different bargaining settings. *Journal of Labor Economics*, 23(4), 875-902.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, 32(2), 189-218.

Chen, Y. C., Genovese, C. R., Tibshirani, R. J., & Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, 44(2), 489-514.

Combes, P.P., G. Duranton, and L.Gobillon (2008). Spatial Wage Disparities: Sorting Matters!. *Journal of Urban Economics* 63(2), 723-742.

Combes, Pierre-Philippe, et al. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica* 80.6: 2543-2594.

Combes, P.P., and L. Gobillon (2015). The Empirics of Agglomeration Economies. *in* G. Duranton, J. V. Henderson and W. Strange (eds), *Handbook in Regional and Urban Economics*, Volume 5, Amsterdam (Holland), Elsevier Press.

Costa, Dora L. and Matthew E. Kahn (2000), "Power Couples: Changes in the Locational Choice of the College Educated, 1940-1990," *Quarterly Journal of Economics*, Volume CXV, 1287-1315.

De la Roca, Jorge (2017). Selection in initial and return migration: Evidence from moves across Spanish cities. *Journal of Urban Economics*, 100, 33-53.

De la Roca, Jorge, and Diego Puga (2017). Learning by Working in Big Cities. *The Review of Economic Studies,* 84.1: 106-142.

Drennan, M. P., and H.F. Kelly (2012). Measuring Urban Agglomeration Economies with Office Rents. *Journal of Economic Geography* 12(3),481-507.

Duranton, Gilles, and Diego Puga (2001). Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*: 1454-1477.

Gaubert, Cecile. (forthcoming) Firm sorting and agglomeration. *American Economic Review.*

Glaeser, Edward L., and David C. Mare. (2001) Cities and skills. *Journal of Labor Economics* 19.2: 316-342.

Heim, B. T. (2007). The incredible shrinking elasticities married female labor supply, 1978–2002. *Journal of Human resources*, 42(4), 881-918.

Huang, M., Li, R., & Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503), 929-941.

Kemp, G. C., & Silva, J. S. (2012). Regression towards the mode. *Journal of Econometrics*, 170(1), 92-101.

Lee, Myoung-jae (1989). Mode regression. *Journal of Econometrics* 42.1-3: 337-349.

Lee, M. J. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1-3), 1-19.

Lewbel, A. (2007). Endogenous selection or treatment model estimation. *Journal of Econometrics*, 141(2), 777-806.

Saumard, A., & Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8, 45: 58-59

Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. *Integrated Public Use Microdata Series: Version 6.0* (dataset). Minneapolis: University of Minnesota, 2015.

Roy, Andrew Donald (1951). Some thoughts on the distribution of earnings. Oxford economic papers 3.2: 135-146.

Rosenthal, S.S and W.C. Strange (2004). Evidence on the Nature and Sources of Agglomeration Economies. in Henderson, J.V. and Thisse, J.-F. (Eds.), *Handbook of Urban and Regional Economics, Volume 4*, Amsterdam: Elsevier, 2129-2172.

Rosenthal, S. S., and W. C. Strange (2008). The Attenuation of Human Capital Spillovers. *Journal of Urban Economics* 64(2), 373-389.

Rosenthal, Stuart S., and William C. Strange (2012). Female Entrepreneurship, Agglomeration, and a New Spatial Mismatch. *Review of Economics and Statistics*, 94(3), 764-788.

Yao, W., & Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3), 656-67

Table 1: Summary statistics of the individual-level data

Panel A
Sales per worker (controlling for type of law firms) for law establishments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	5th quantile	25th quantile	50th quantile	75th quantile	95th quantile	mean	Observation
All Firms	46,666	59,000	67,609	82,222	121,446	72,639	545,873
Young Firms (<= 5 years)	38,422	52,287	60,000	70,000	96,531	62,735	74,079
Old Firms (> 5 years)	48,255	60,000	70,000	84,261	123,541	74,194	471,794

Panel B
Adjust wage for married non-Hispanic white female full-time workers, age 25-54

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	5th quantile	25th quantile	50th quantile	75th quantile	95th quantile	mean	Observation
College degree or more	10.27	17.85	22.66	28.09	41.95	24.08	152,704
High school degree or less	9.01	12.51	15.01	18.10	25.17	15.75	153,168

Panel C
Adjust wage summary statistics for male non-Hispanic white full-time workers, age 25-54

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	5th quantile	25th quantile	50th quantile	75th quantile	95th quantile	mean	Observation
College degree or more	6.59	21.17	29.08	37.52	81.18	32.49	383,728
High school degree or less	6.83	12.74	15.38	19.75	29.58	16.40	393,598

Note: Law firm data are from Dun and Bradstreet for December 2016. The sample is restricted to single-site firms which excludes roughly 2 percent of establishments. MSAs are restricted to those with 100,000 or more population that have at least 30 or more law firms present for both young and old classifications of law firms. Married female individual-level data are obtained from the 2000 Census. Hourly wage is adjusted by controlling for age, education, occupation and industry fixed effects. The sample is restricted to cities with at least 100,000 or more population that have at least 100 or more observation in each education category.

Table 2: Summary statistics of the mode estimates across MSAs

Panel A
Modal sales per worker estimates for law establishments

	(1)	(2)	(3)	(4)	(5)	(6)
	Min	Max	Median	Mean	Std.	Observation
All Firms	50,000	70,000	55,000	57,562	4,383	239
Young Firms (<= 5 years)	45,000	65,000	55,000	55,659	3,781	239
Old Firms (> 5 years)	50,000	75,000	55,000	58,075	4,867	239

Panel B
Modal wage estimates for married white non-Hispanic full-time female workers, aged 25-54

	(1)	(2)	(3)	(4)	(5)	(6)
	Min	Max	Median	Mean	Std.	Observation
College degree or more	15.00	27.00	21.00	21.27	2.29	216
High school degree or less	12.00	18.00	15.00	14.03	1.56	216

Panel C
Modal adjusted wage summary statistics for male non-Hispanic white full-time worker, age 25-54

	(1)	(2)	(3)	(4)	(5)	(6)
	Min	Max	Median	Mean	Std.	Observation
College degree or more	21.00	39.00	27.00	27.52	2.98	262
High school degree or less	9.00	21.00	15.00	13.78	1.79	262

Note: Law firm data are from Dun and Bradstreet for December 2016. The sample is restricted to single-site firms which excludes roughly 2 percent of establishments. MSAs are restricted to those with 100,000 or more population that have at least 30 or more law firms present for both young and old classifications of law firms. Individual-level data are obtained from the 2000 Census. Wage is adjusted by controlling for occupation, industry, age and education fixed effects. The sample is restricted to cities with at least 100,000 or more population that have at least 100 or more observation in each education category. Bandwidth used to define modal wage is $3. Bandwidth used to define modal sales per worker is $5,000.

Table 3: OLS results based on law establishments

	(1)	(2)	(3)	(4)
Panel A: All Firms	CDF of Sale/Worker evaluated at the Mode	Log(Sale/Worker) at the Mode	Log(Sale/Worker) at the Mean	Coefficient difference (3) - (2)
Log population in MSA	-0.0210	0.0105	0.0169	0.0064
	(-3.81)	(2.20)	(7.23)	(1.68)
R-squared	0.058	0.021	0.164	0.012
Observations	239	239	239	239
Panel B: Young Firms (<= 5 years)				
Log population in MSA	0.0158	0.0305	0.0174	-0.0131
	(2.81)	(9.75)	(6.44)	(-3.94)
R-squared	0.022	0.224	0.003	0.041
Observations	239	239	239	239
Panel C: Old Firms (> 5 years)				
Log population in MSA	-0.0226	0.0084	0.0163	0.0080
	(-3.28)	(1.54)	(6.41)	(1.69)
R-squared	0.054	0.012	0.132	0.154
Observations	239	239	239	239

Note: T-ratios based on robust standard errors in parentheses. Data are from Dun and Bradstreet for December 2016. The sample is restricted to single-site firms which excludes roughly 2 percent of establishments. MSAs are restricted to those with 100,000 or more population that have at least 30 or more law firms present for both young and old classifications of law firms.

Table 4: OLS results based on married female white full-time workers, age 25-54

	(1) CDF of wage evaluated at the Mode	(2) Log(wage) at the Mode	(3) Log(wage) at the Mean	(4) Coefficient difference (3) - (2)
Panel A: College degree or more				
Log population in MSA	-0.0195	0.0232	0.0428	0.0197
	(-3.02)	(3.60)	(9.98)	(3.43)
R-squared	0.033	0.049	0.312	0.043
Observations	216	216	216	216
Panel B: High school degree or less				
Log population in MSA	-0.0074	0.0377	0.0388	0.0012
	(-1.17)	(6.50)	(10.03)	(0.24)
R-squared	0.005	0.126	0.291	0.000
Observations	216	216	216	216

Note: T-ratios based on robust standard errors in parentheses. Married female worker data are obtained from the 2000 Census. Wage is adjusted by controlling for occupation, industry, age and education fixed effects. The sample is restricted to cities with at least 100,000 or more population that have at least 100 or more observation in each education category.

Table 5: OLS results based on male white non-Hispanic full-time workers, age 25-54

Panel A: College degree or more	(1) CDF of wage evaluated at the Mode	(2) Log(wage) at the Mode	(3) Log(wage) at the Mean	(4) Coefficient difference (3) - (2)
Log population in MSA	-0.0082	0.0245	0.0452	0.0207
	(-1.79)	(4.70)	(12.58)	(3.87)
R-squared	0.009	0.055	0.352	0.041
Observations	262	262	262	262
Panel B: High school degree or less				
Log population in MSA	0.0014	0.0440	0.0361	-0.0079
	(0.34)	(6.88)	(7.46)	(-1.75)
R-squared	0.000	0.125	0.173	0.009
Observations	262	262	262	262

Note: T-ratios based on robust standard errors in parentheses. Male worker data are obtained from the 2000 Census. Wage is adjusted by controlling for occupation, industry, age and education fixed effects. The sample is restricted to cities with at least 100,000 or more population that have at least 100 or more observation in each education category.

Figure 1: Productivity Distributions with Agglomeration Economies

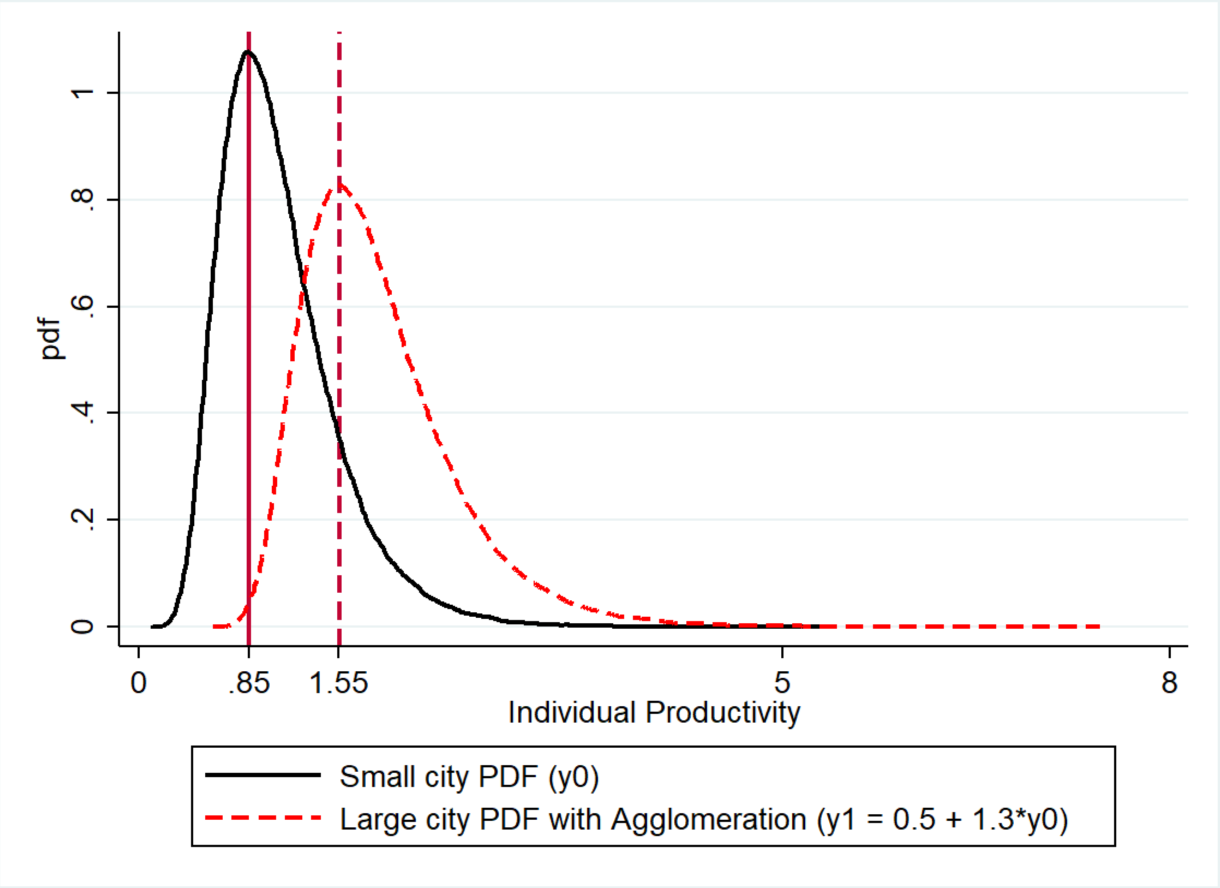
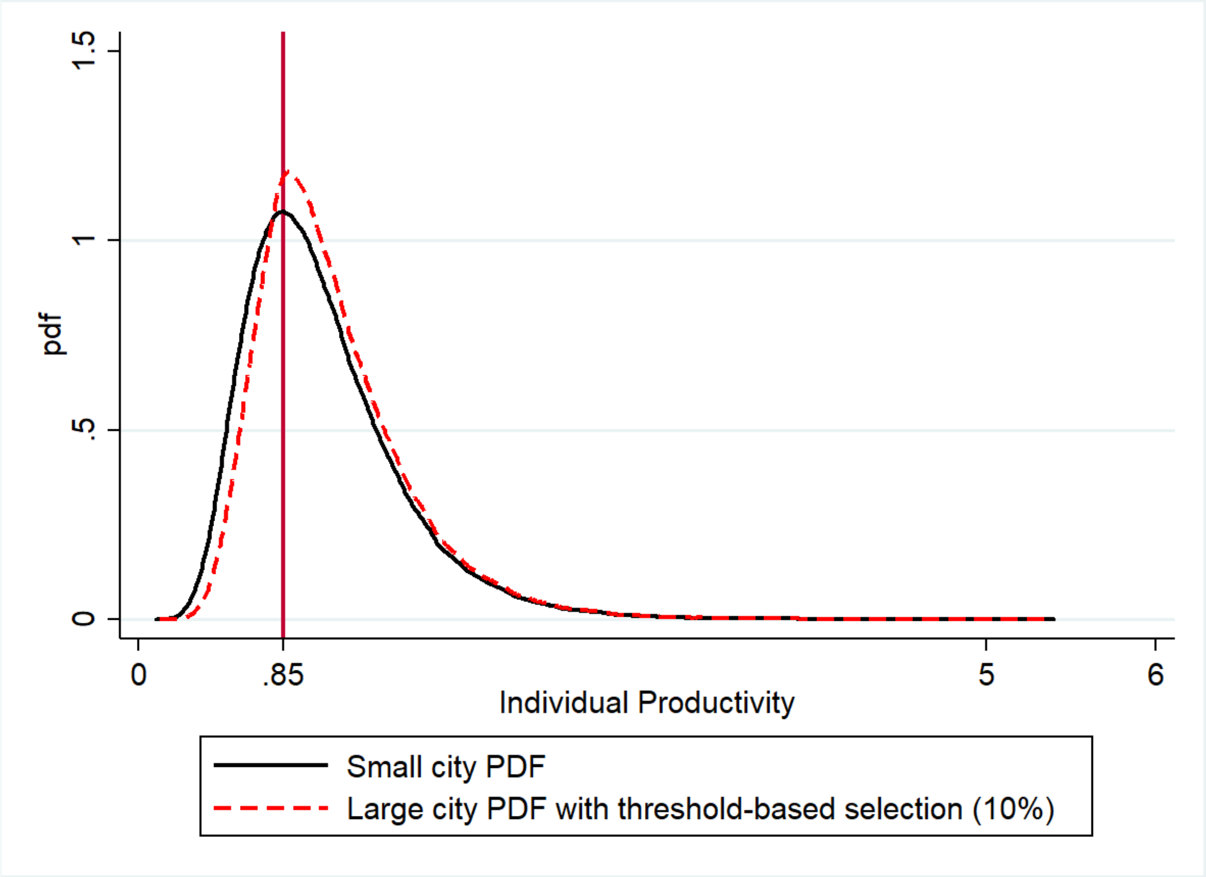


Figure 2: Productivity Distributions with Threshold Effects



Note: This figure illustrates a case where 10% of the workers are selected out of the large city's labor market.

Figure 3: Productivity Distributions with Agglomeration Economies and Threshold Effects

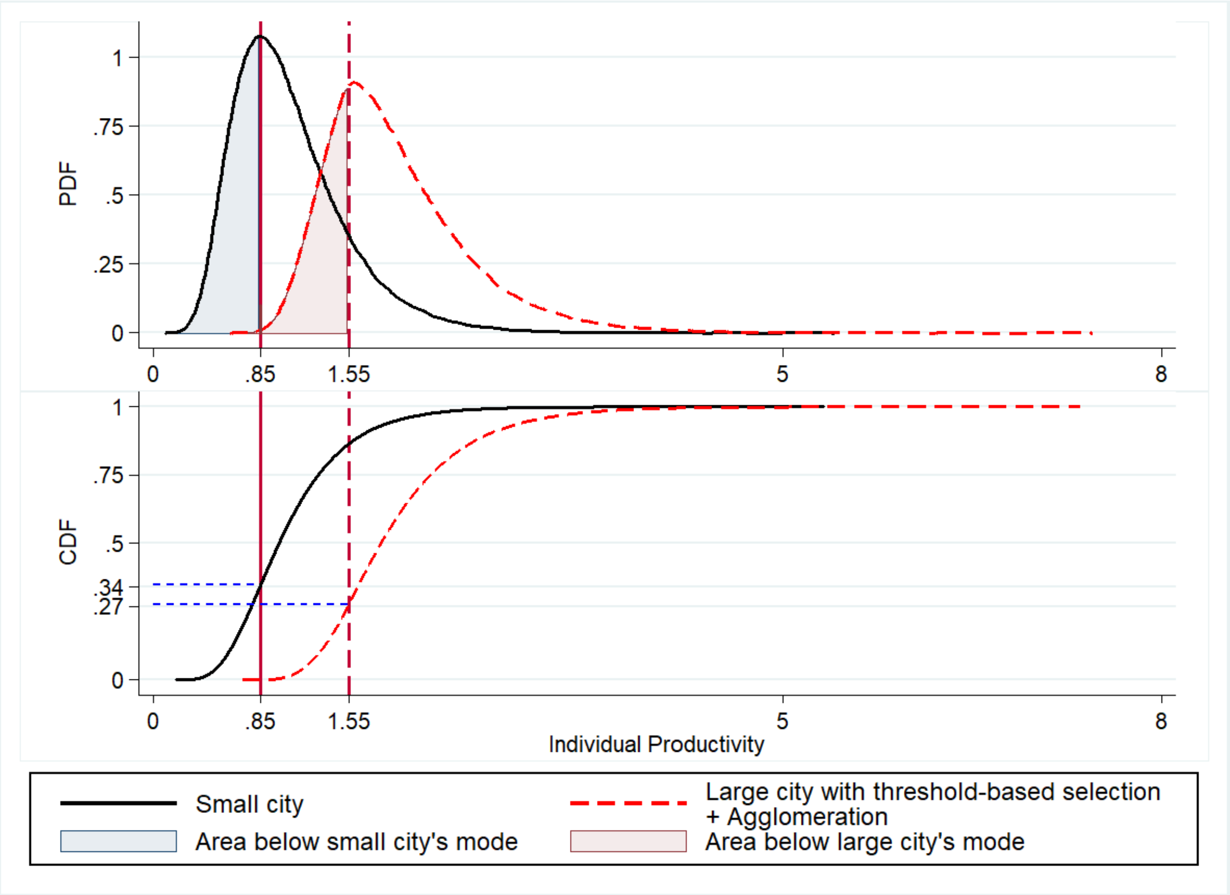
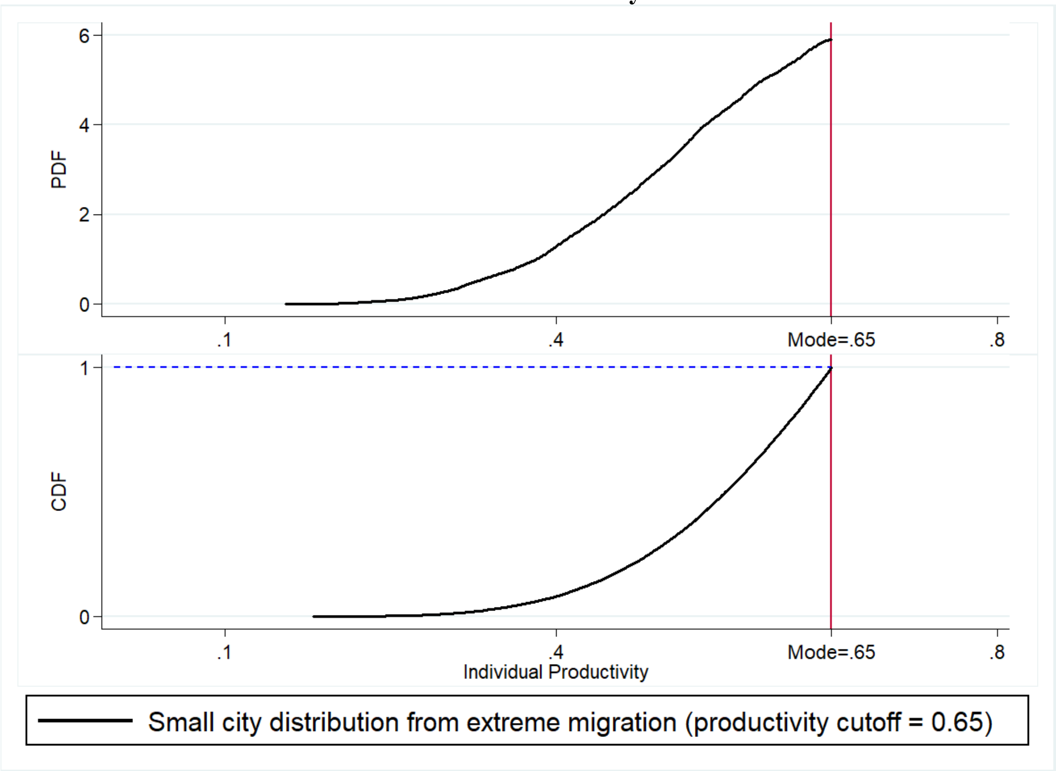


Figure 4a: Small and Large City Productivity Distributions from Extreme Migration Effects (Original Mode=0.85 and Migration Productivity Cutoff = 0.65)

Panel A: Small City



Panel B: Large City

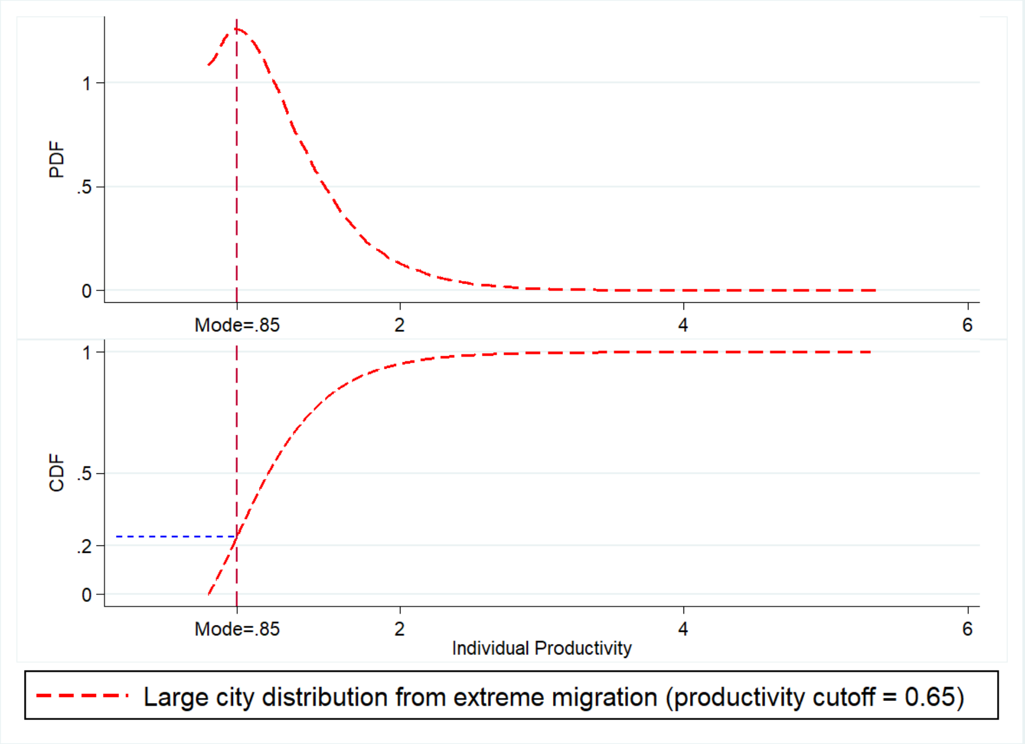
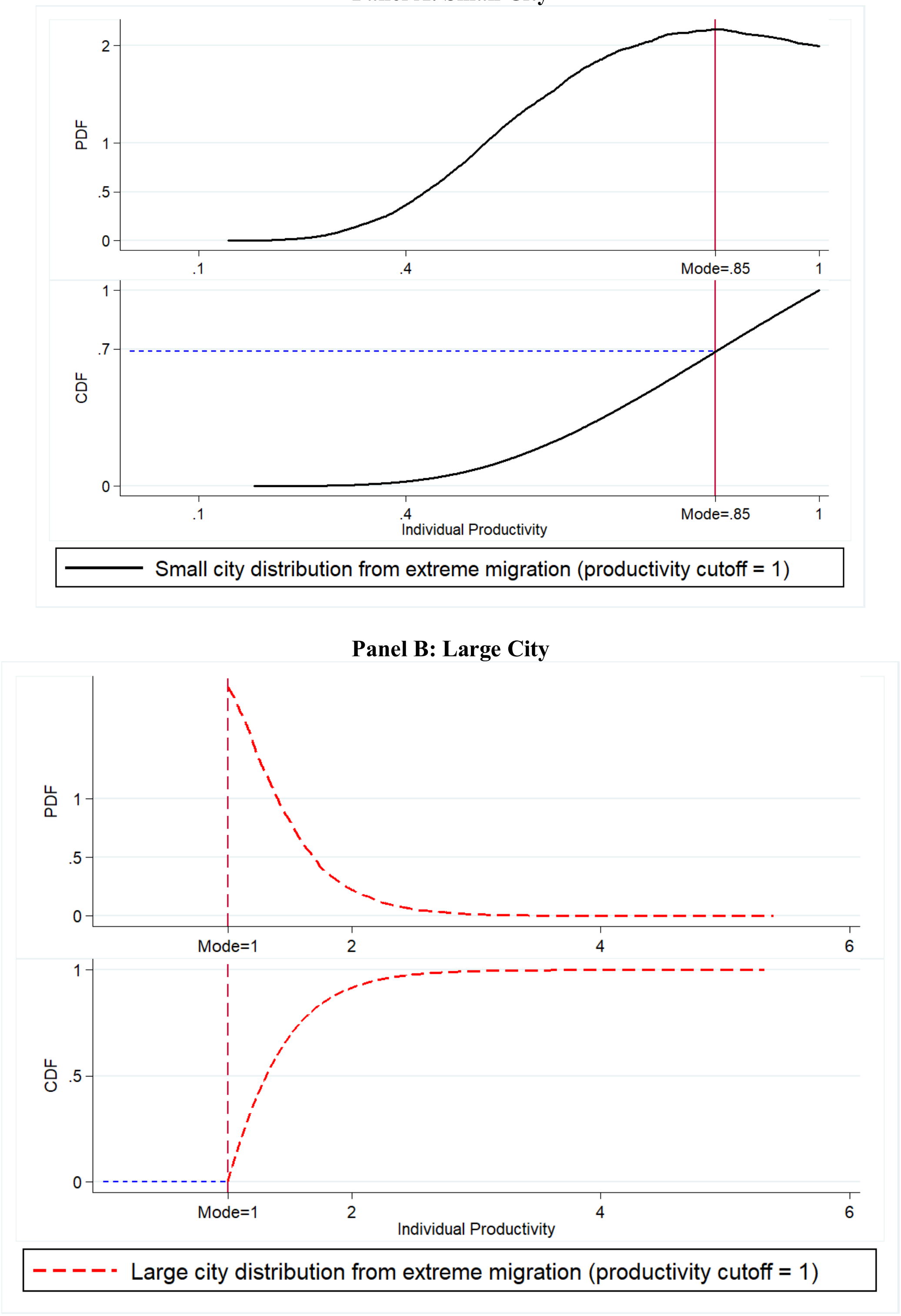


Figure 4b: Small and Large City Productivity Distributions from Extreme Migration Effects
(Original Mode=0.85 and Migration Productivity Cutoff = 1)



Panel A: Small City

Panel B: Large City

Figure 5a: Productivity Distributions with Migration Effects

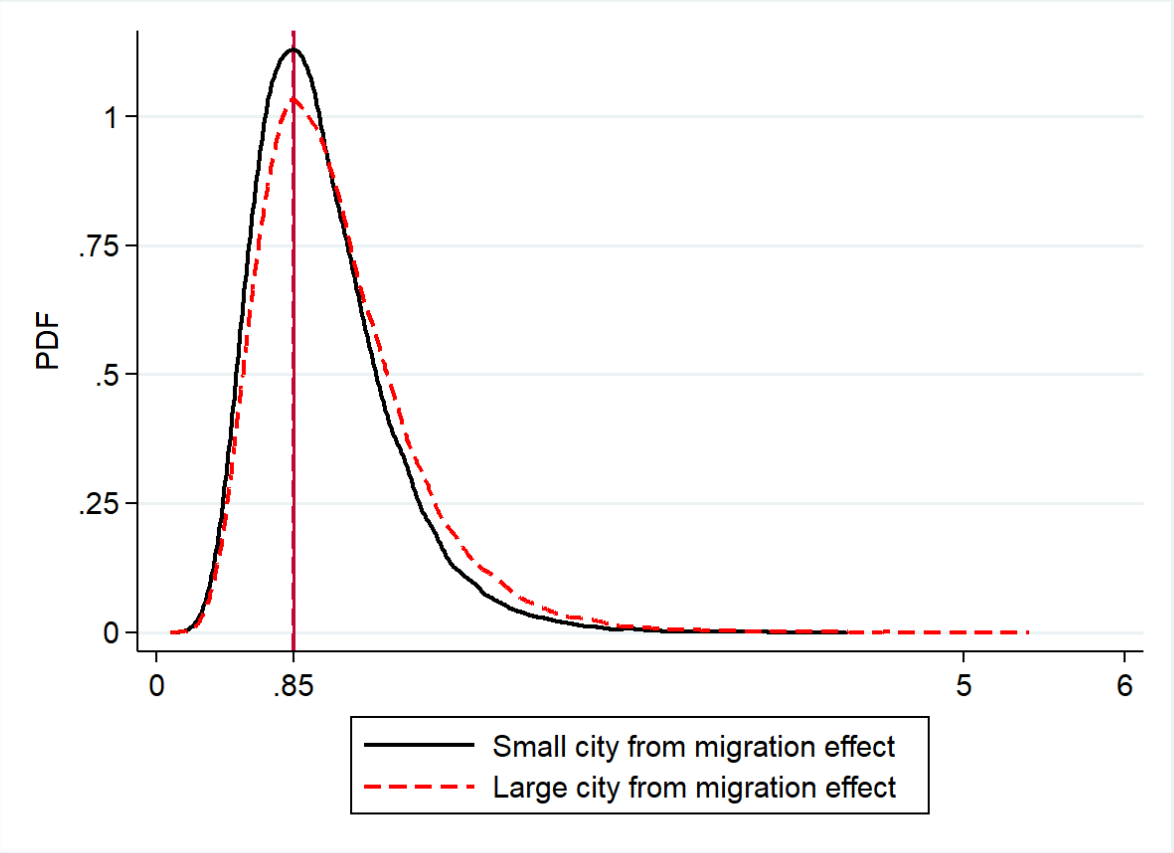


Figure 5b: Productivity Distributions with Agglomeration Economies and Migration Effects

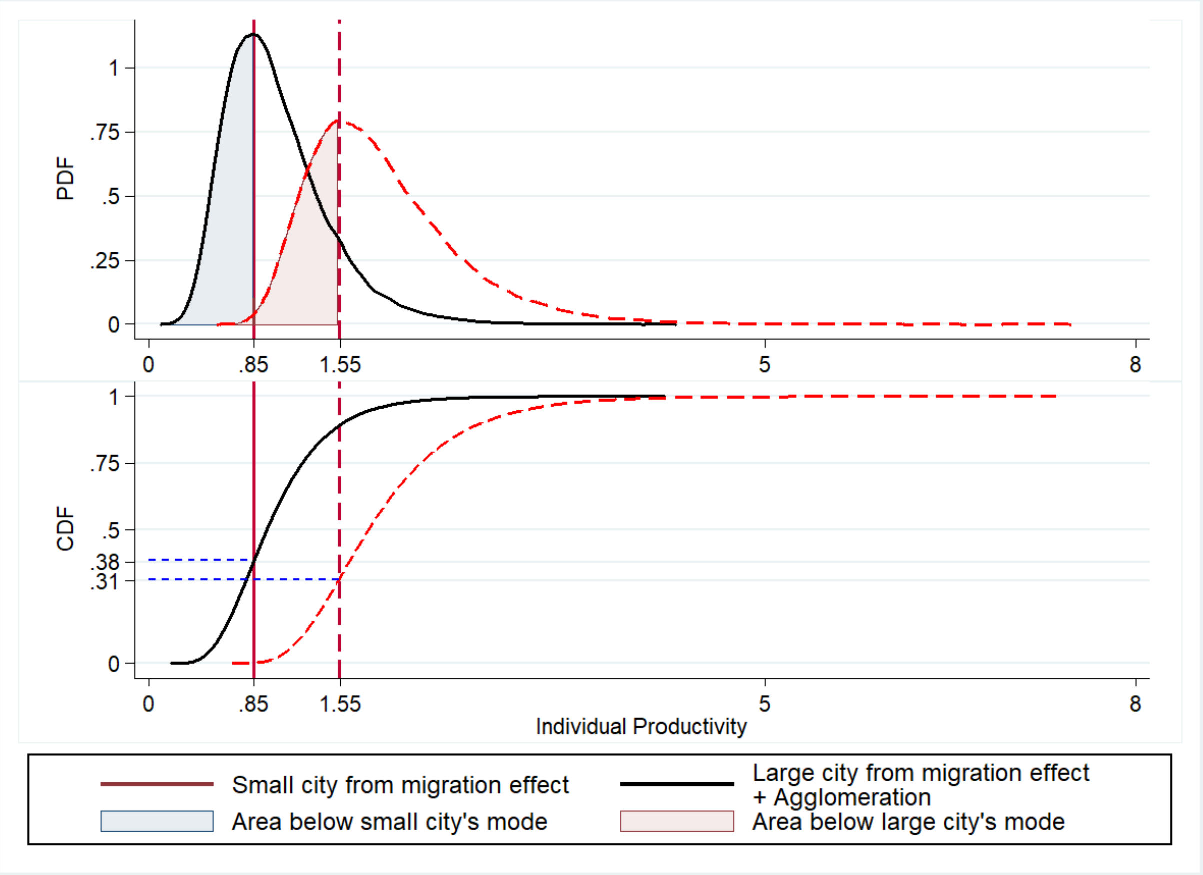


Figure 6: Slope Conditions and Shifts in the Mode

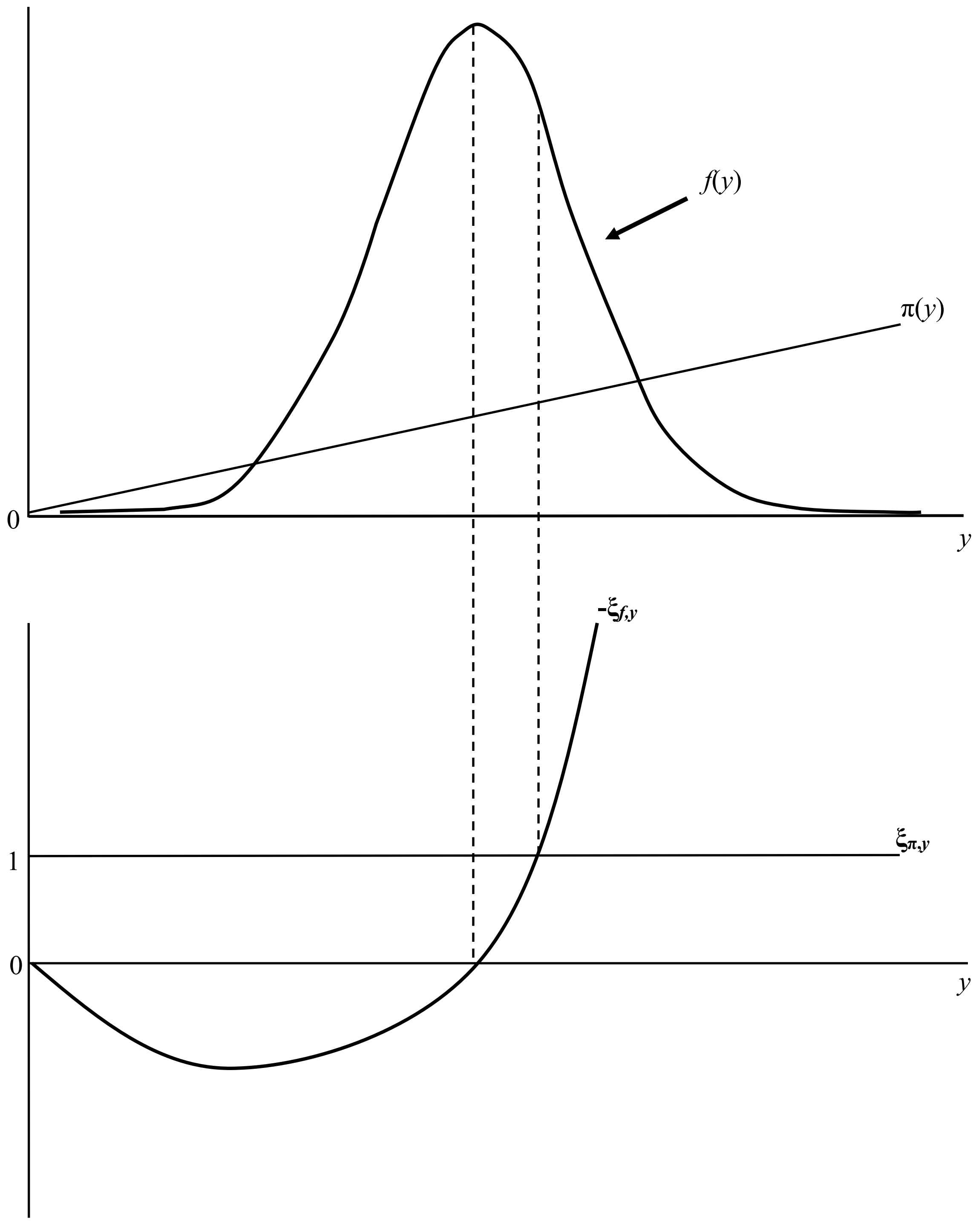
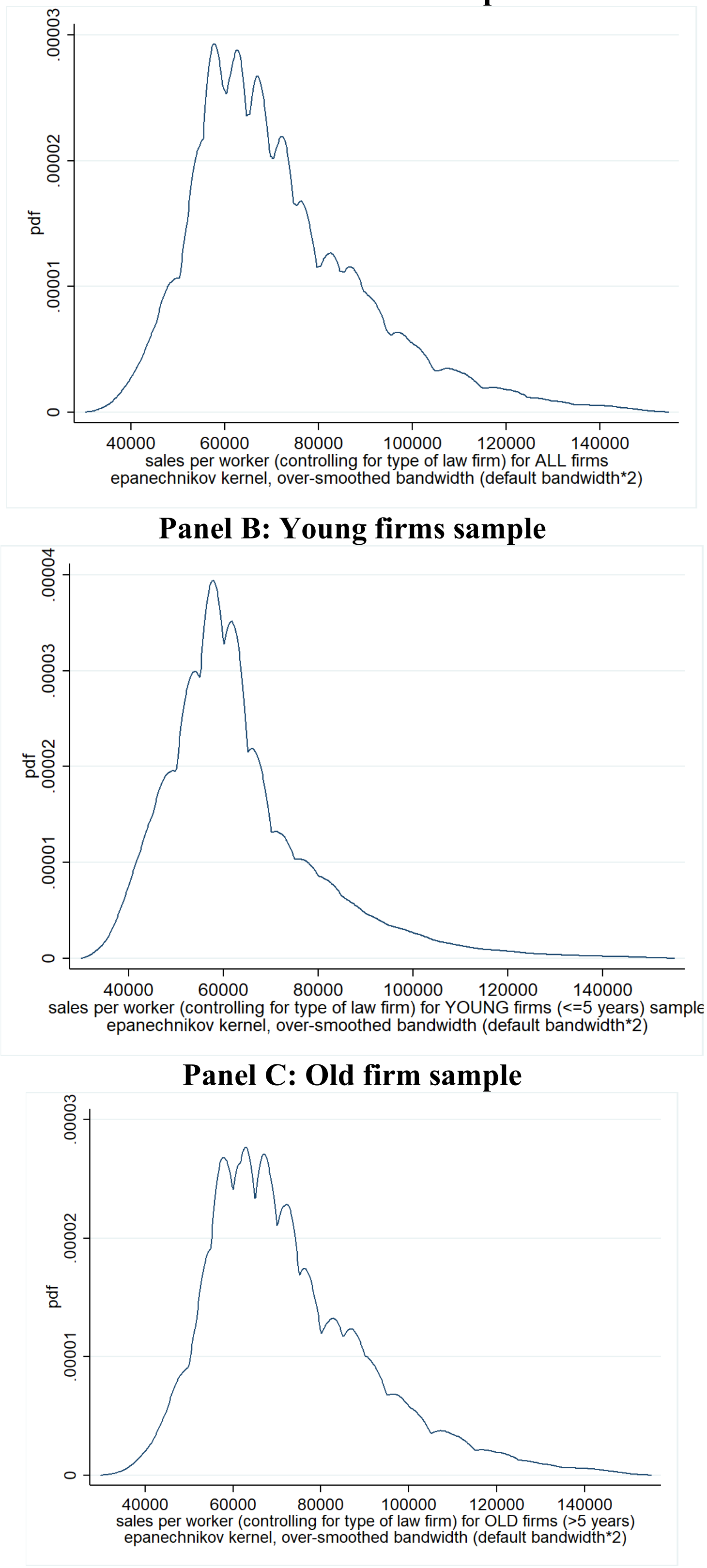


Figure 7a: Sale per worker kernel density estimation for law firms in the Unites States

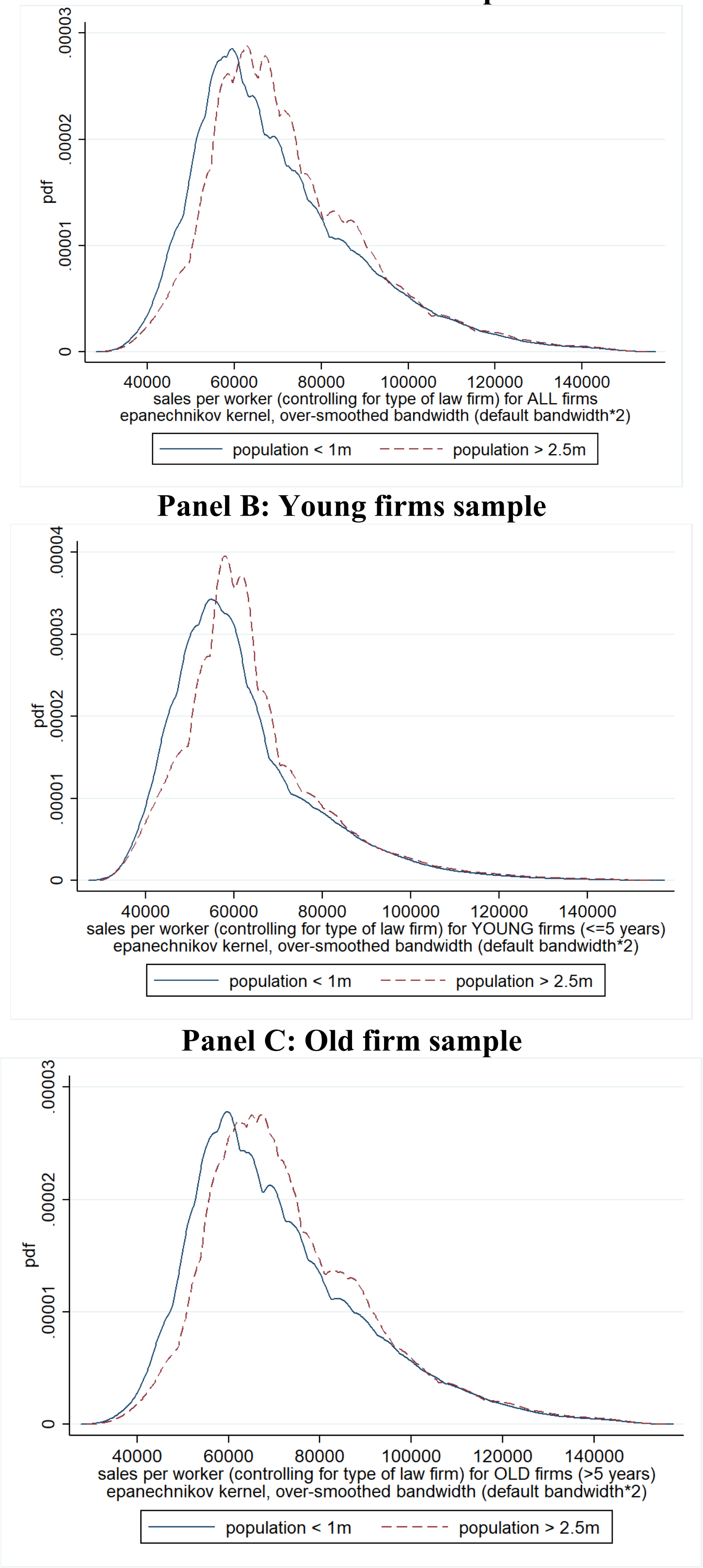


Panel A: All firms sample

Panel B: Young firms sample

Panel C: Old firm sample

Figure7b: Sale per worker kernel density estimation for law firms in small versus large cities

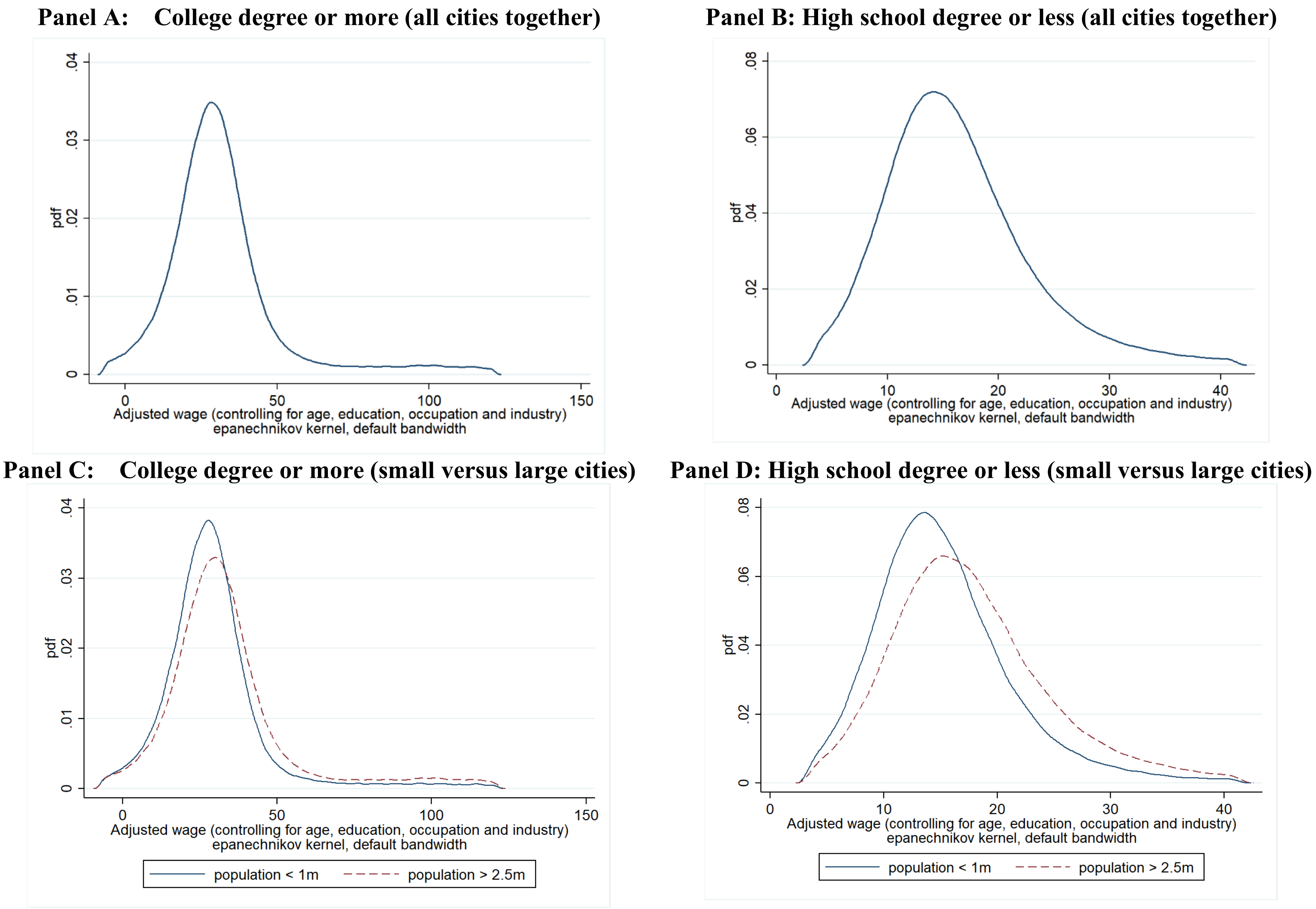


Panel A: All firms sample

Panel B: Young firms sample

Panel C: Old firm sample

Figure 8: Adjusted wage kernel density estimation for male white full-time worker, age 25-54



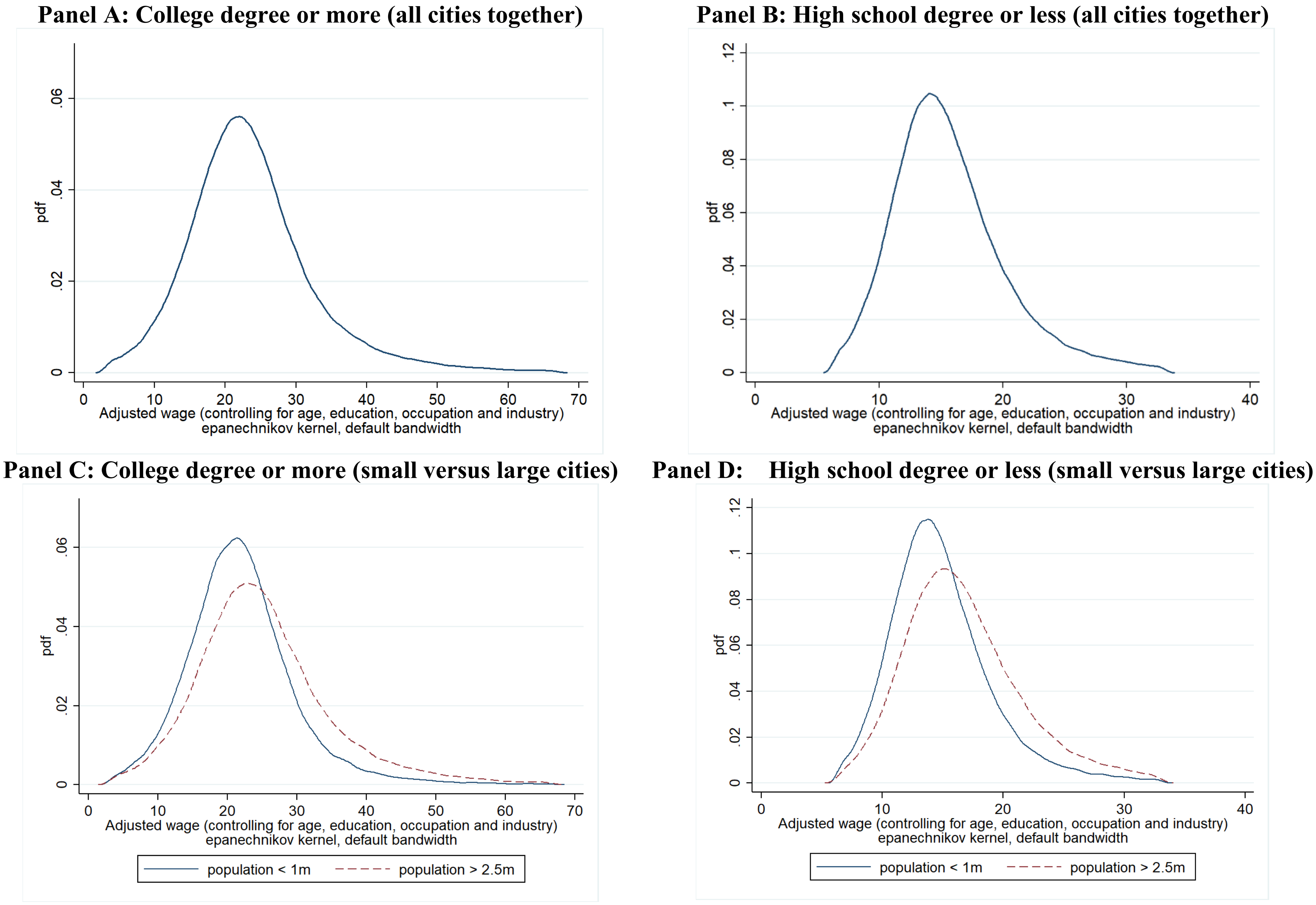
Panel A: College degree or more (all cities together)

Panel B: High school degree or less (all cities together)

Panel C: College degree or more (small versus large cities)

Panel D: High school degree or less (small versus large cities)

Figure 9: Adjusted wage kernel density estimation for married female white full-time workers, age 25-54



Panel A: College degree or more (all cities together)

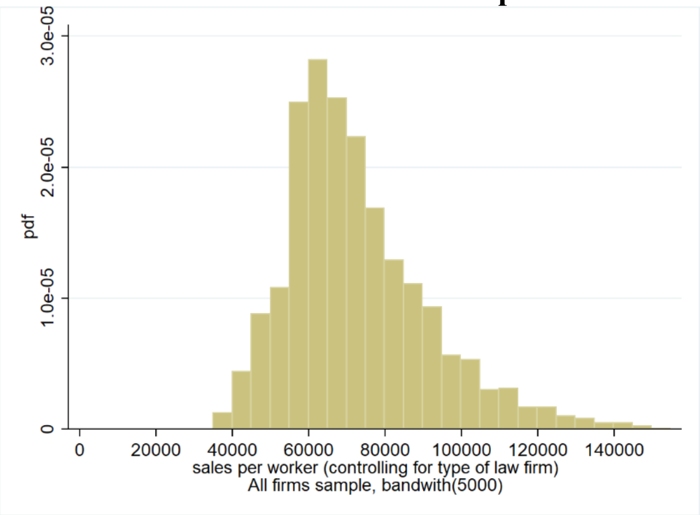
Panel B: High school degree or less (all cities together)

Panel C: College degree or more (small versus large cities)

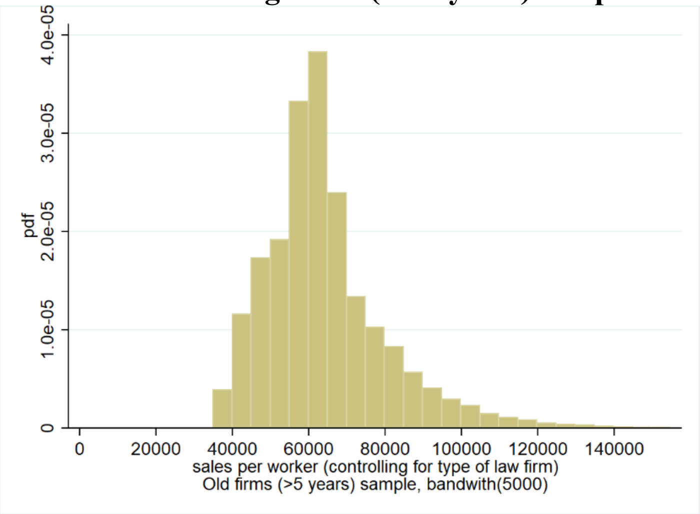
Panel D: High school degree or less (small versus large cities)

Figure 10a: Histogram estimation of sale per worker for law firms (bandwidth $5,000)

Panel A: ALL firms sample



Panel B: Young firms (<= 5 years) sample



Panel C: Old firms (> 5 years) sample

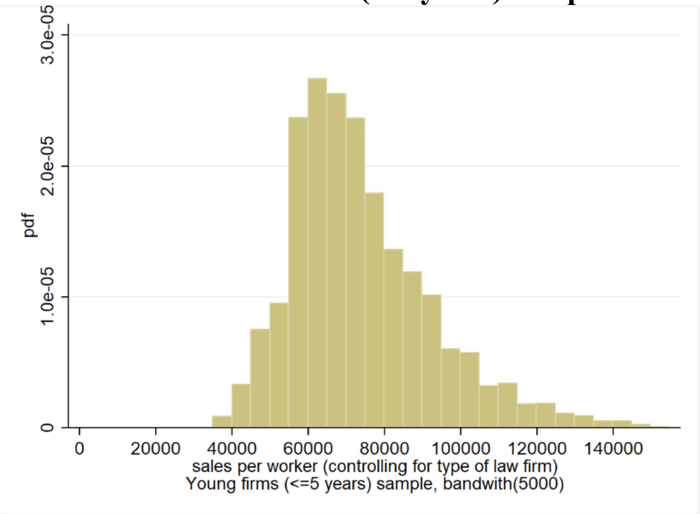
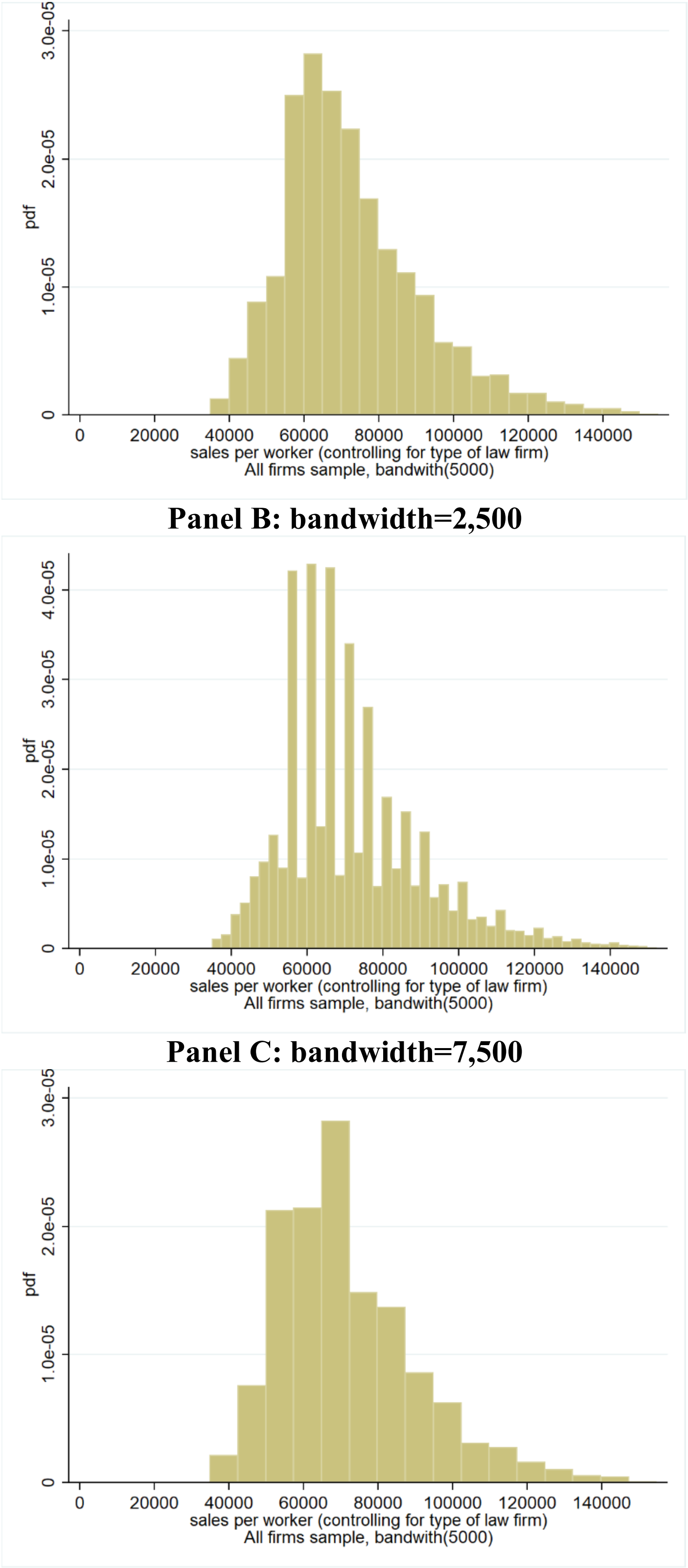


Figure 10b: Histogram estimation of sale per worker for law firm using different bandwidth

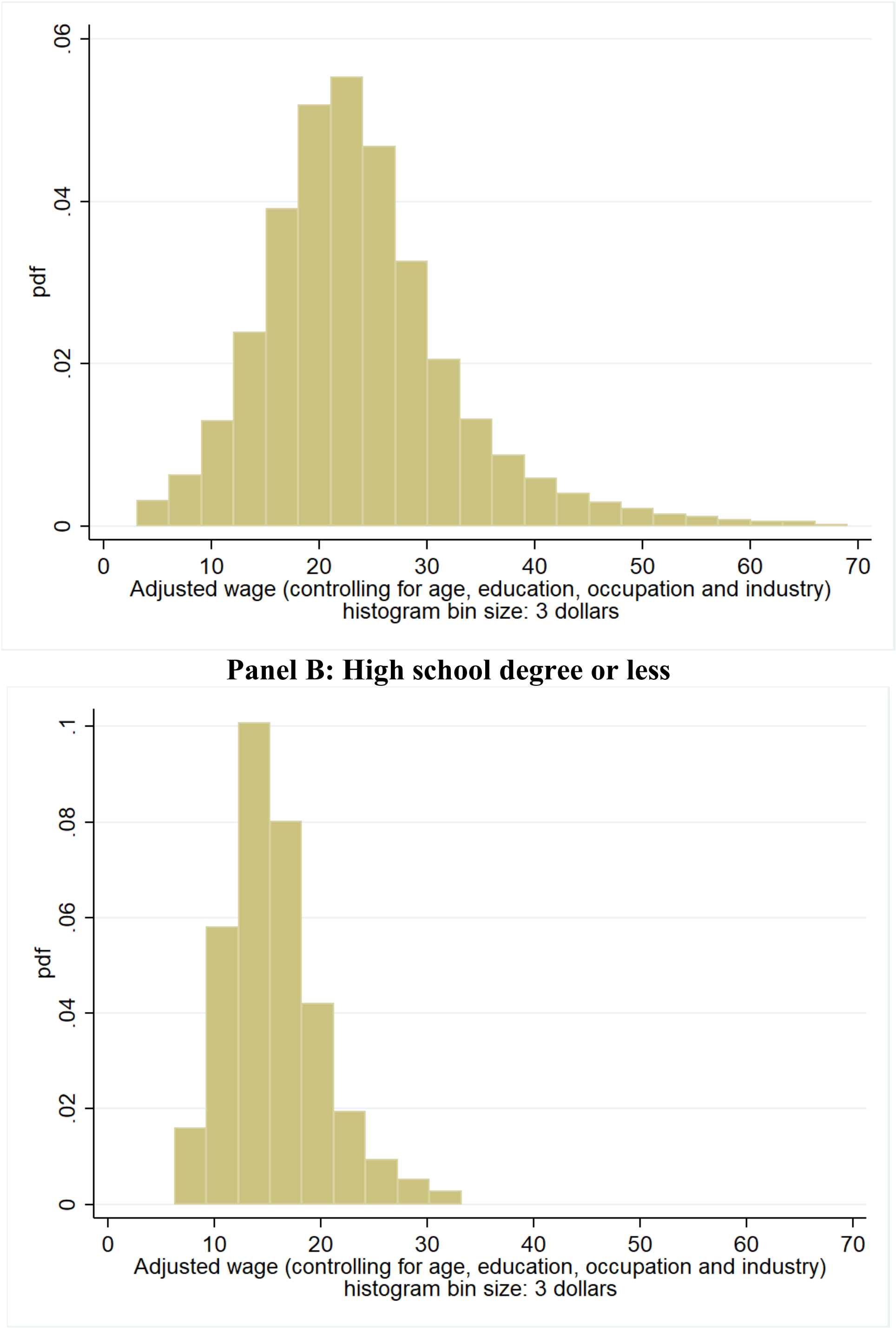


Panel A: bandwidth=5,000

Panel B: bandwidth=2,500

Panel C: bandwidth=7,500

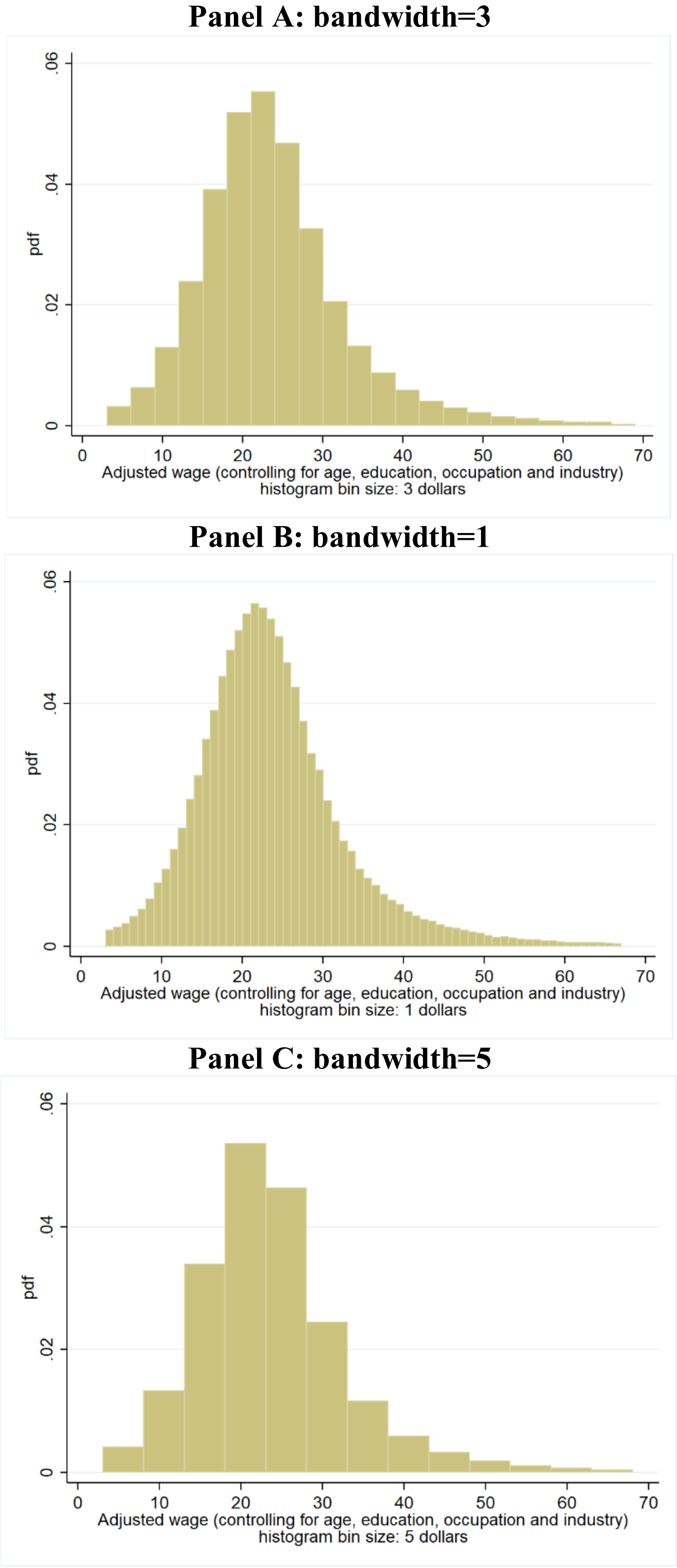
Figure 11a: Histogram estimation of adjusted wage for married female white full-time workers (age 25-54), (bandwidth $3)



Panel A: College degree or more

Panel B: High school degree or less

Figure 11b: Histogram estimation of adjusted wage for skilled (college degree or more) married female white full-time workers (age 25-54) using different bandwidth

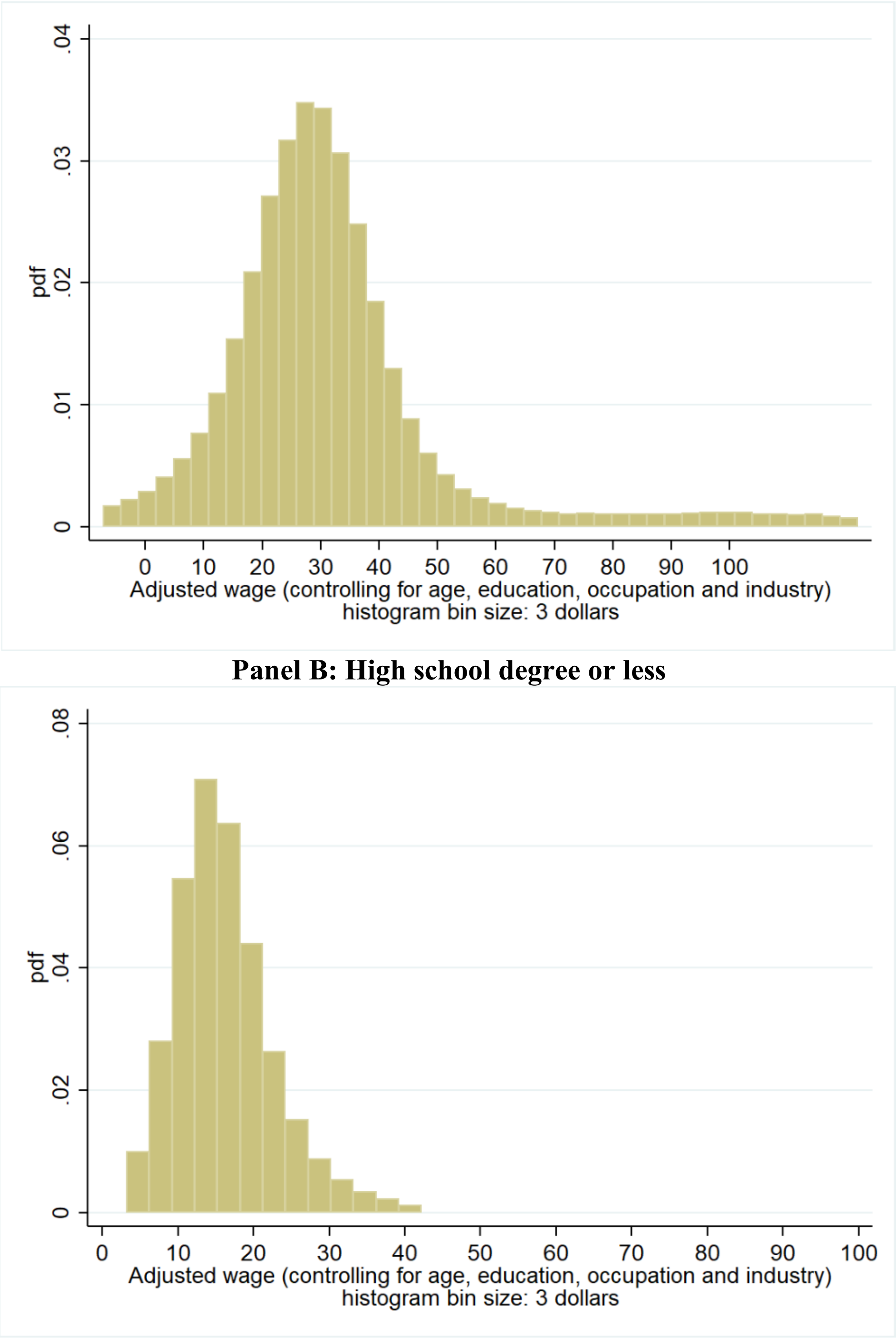


Panel A: bandwidth=3

Panel B: bandwidth=1

Panel C: bandwidth=5

Figure 12a: Histogram estimation of adjusted wage for male white full-time workers

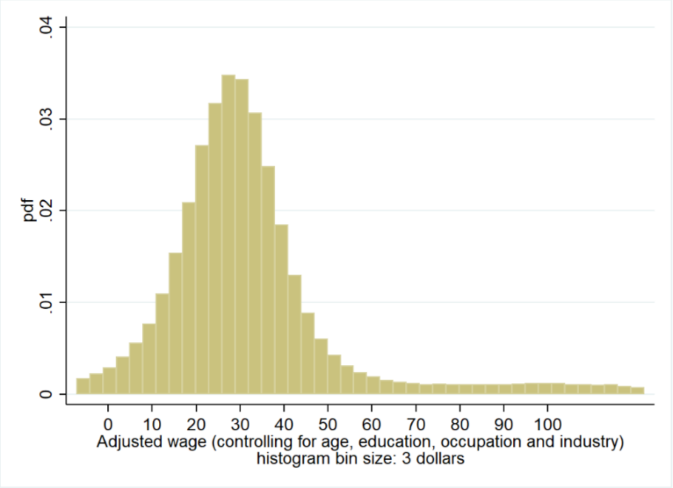


Panel A: College degree or more

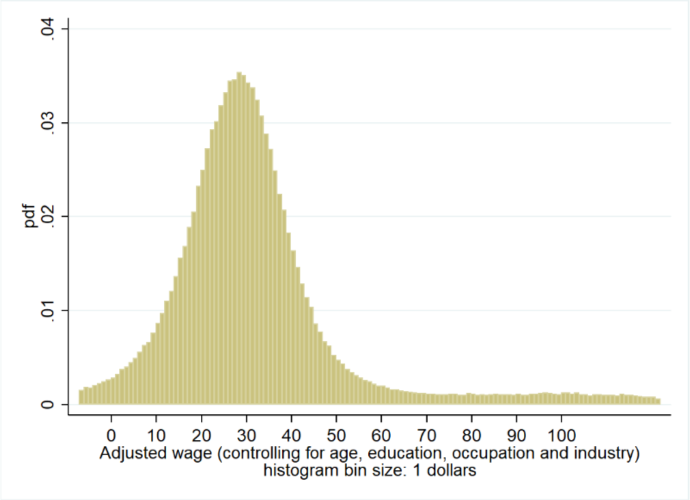
Panel B: High school degree or less

Figure 12b: Histogram estimation of adjusted wage for skilled (college degree or more) male white full-time workers (age 25-54) using different bandwidth

Panel A: bandwidth=3



Panel B: bandwidth=1



Panel C: bandwidth=5

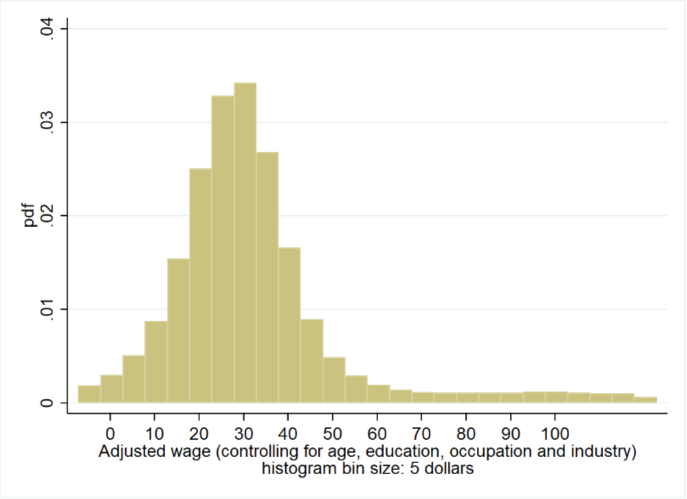
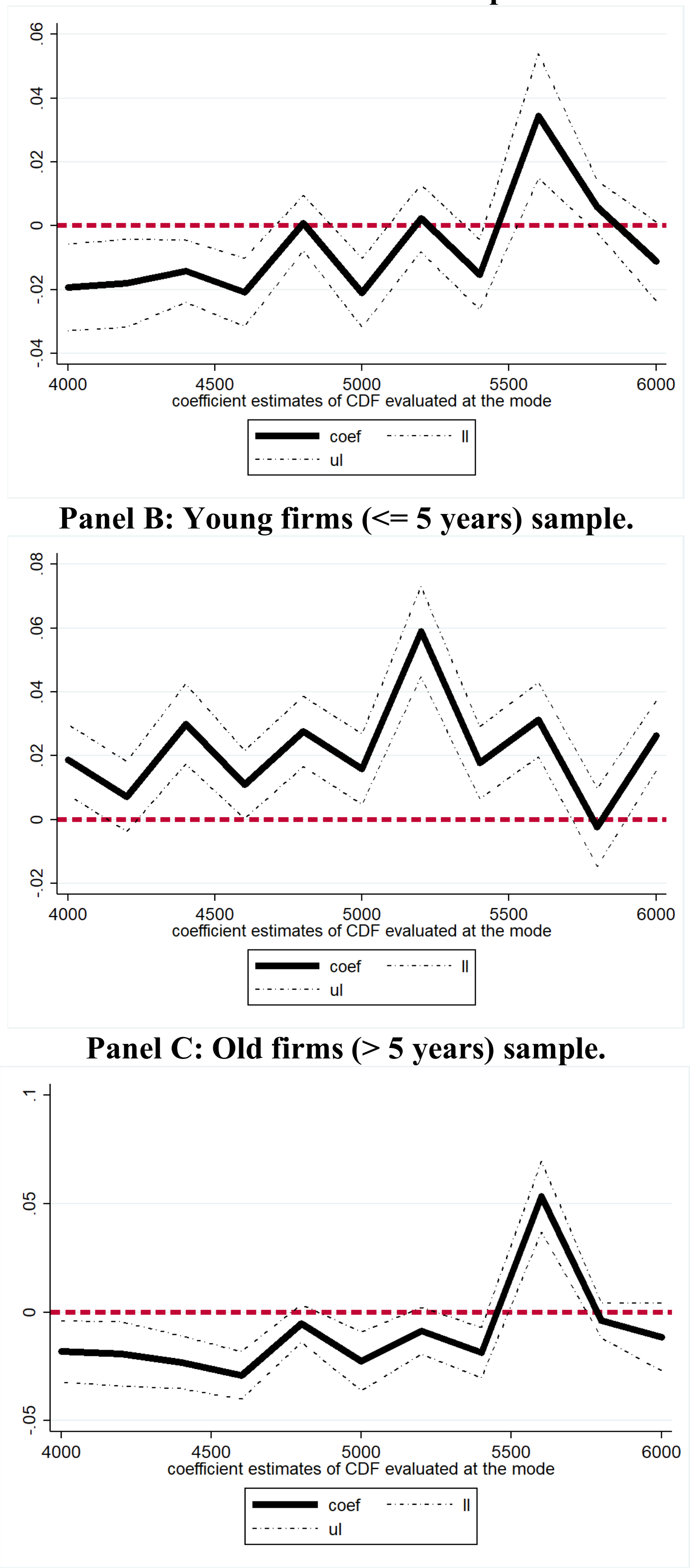


Figure 13a: Law establishment modal CDF robustness check using different bandwidth

Panel A: ALL firms sample.

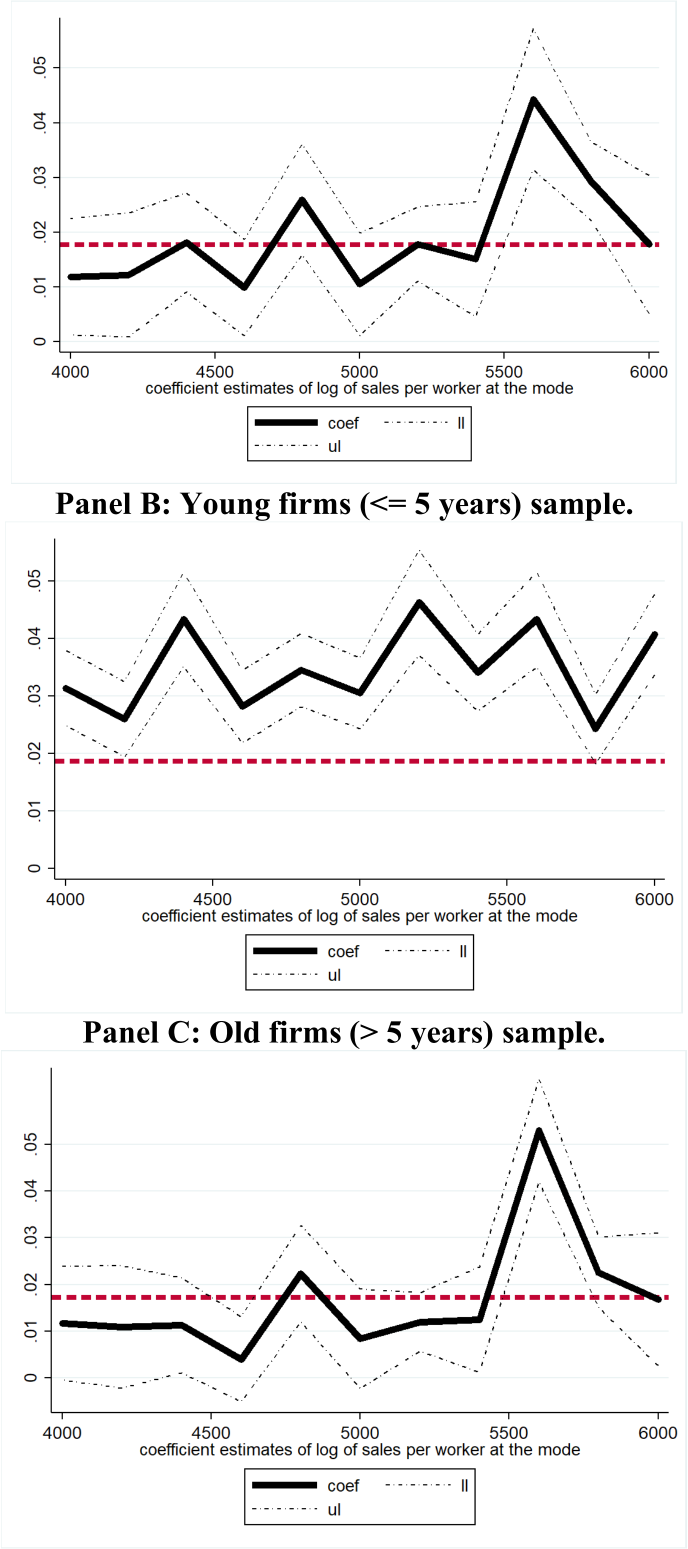


Panel B: Young firms (<= 5 years) sample.

Panel C: Old firms (> 5 years) sample.

Note: Red dashed horizontal line represents value 0

Figure 13b: Law establishment modal return robustness check using different bandwidth



Panel A: ALL firms sample.

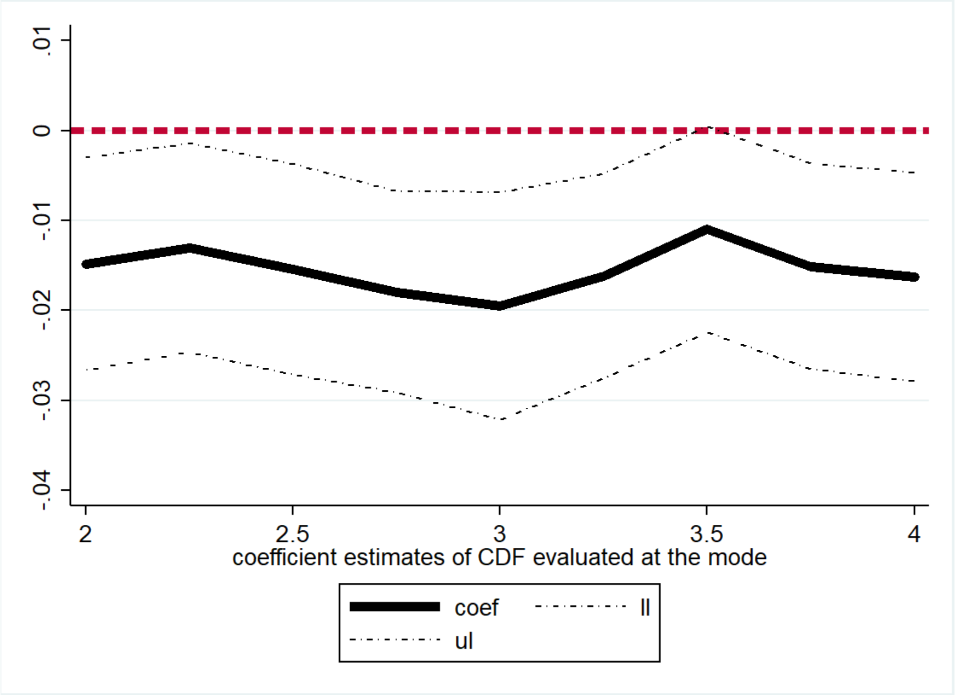
Panel B: Young firms (<= 5 years) sample.

Panel C: Old firms (> 5 years) sample.

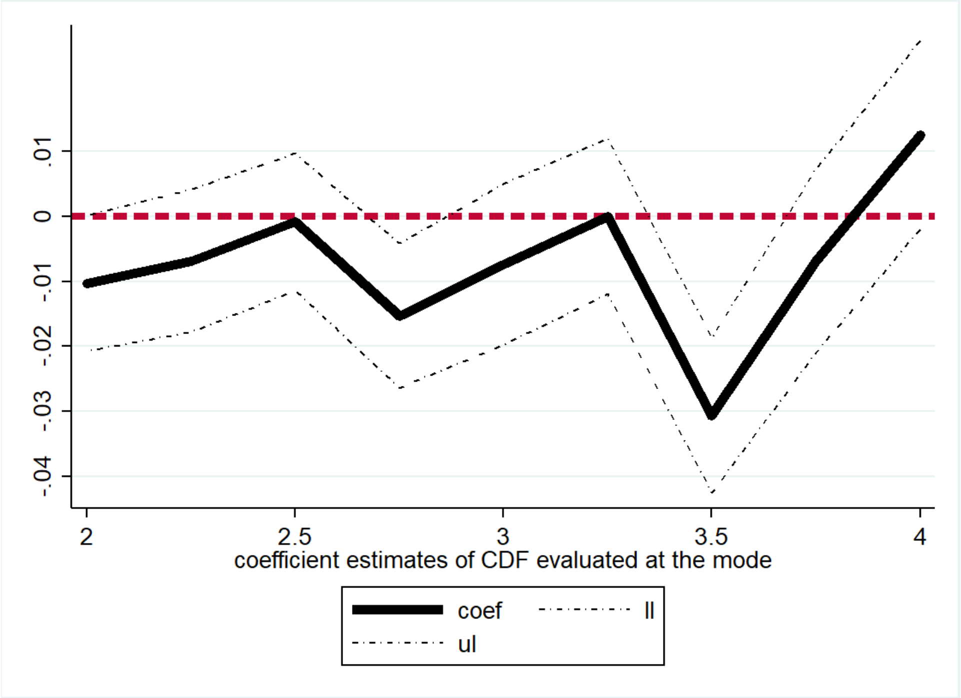
Note: Red dashed horizontal line represents mean return for each sample.

Figure 14a: Married female modal CDF robustness check using different bandwidth

Panel A: College degree or more

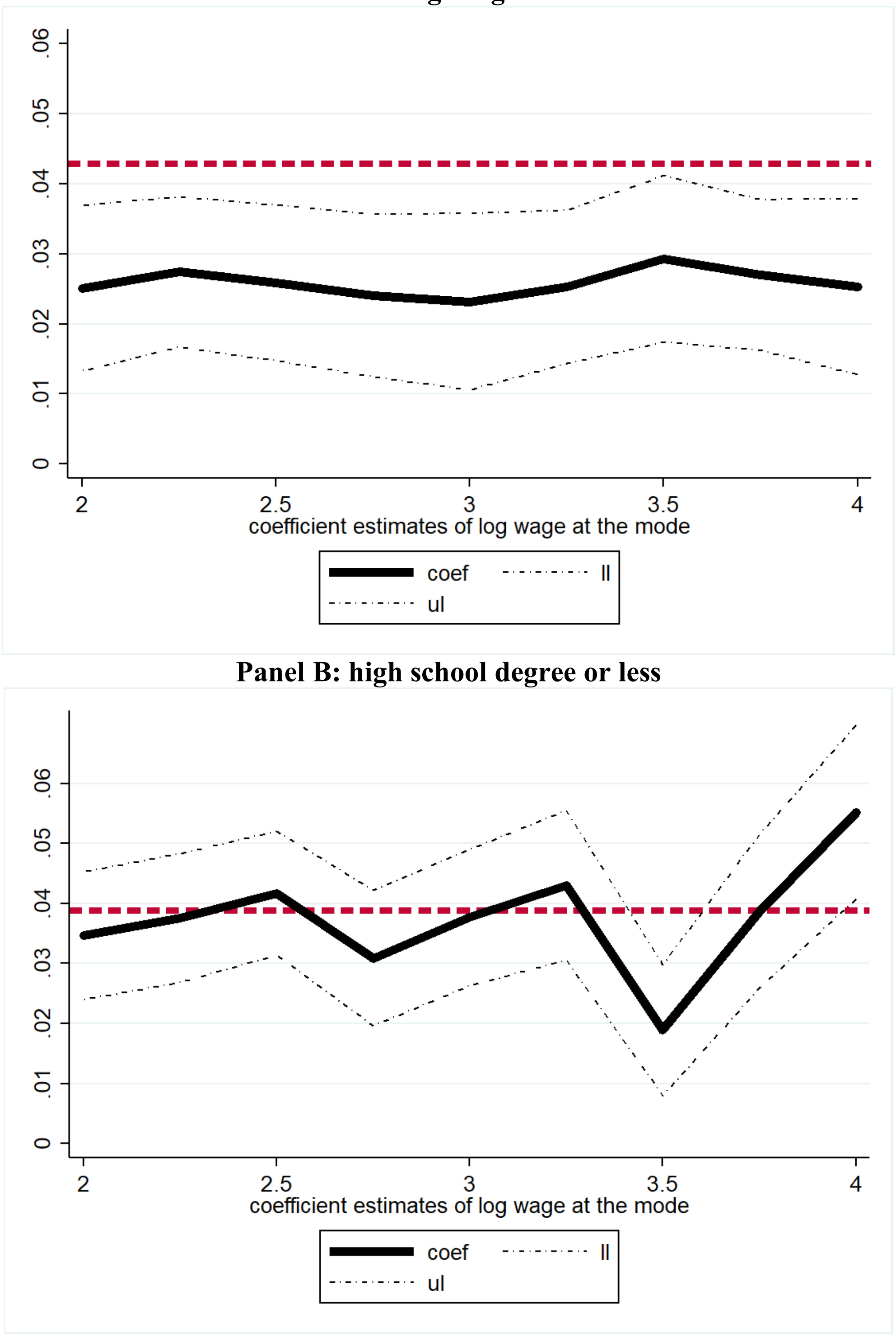


Panel B: high school degree or less



Note: Red dashed horizontal line represents value 0.

Figure 14b: Married female men modal return robustness check using different bandwidth



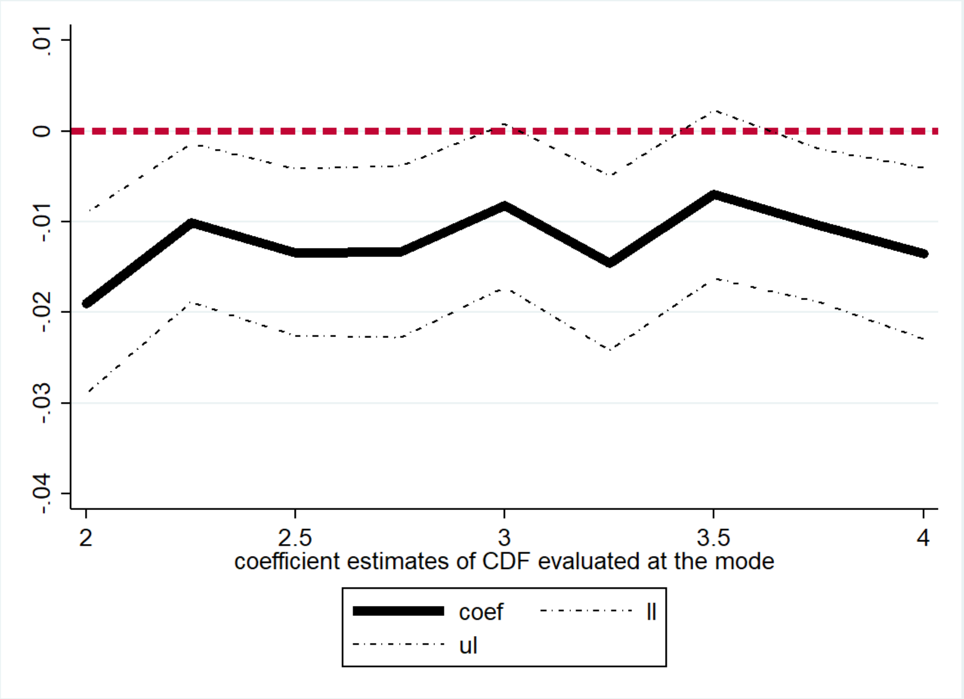
Panel A: College degree or more

Panel B: high school degree or less

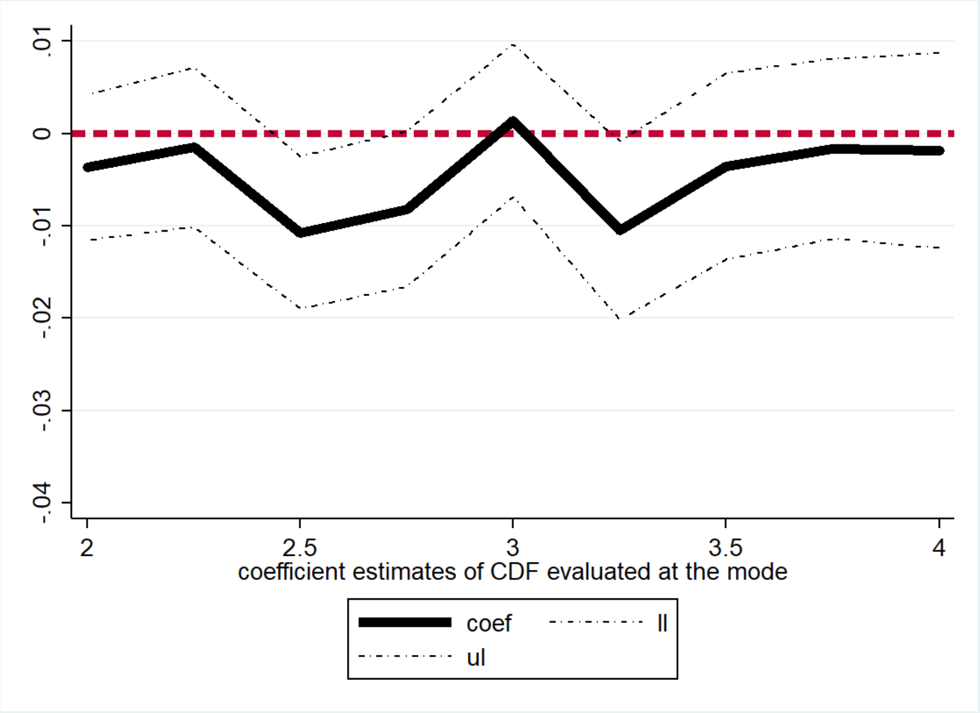
Note: Red dashed horizontal line represents mean return for each sample.

Figure 15a: Male modal CDF robustness check using different bandwidth

Panel A: college degree or more

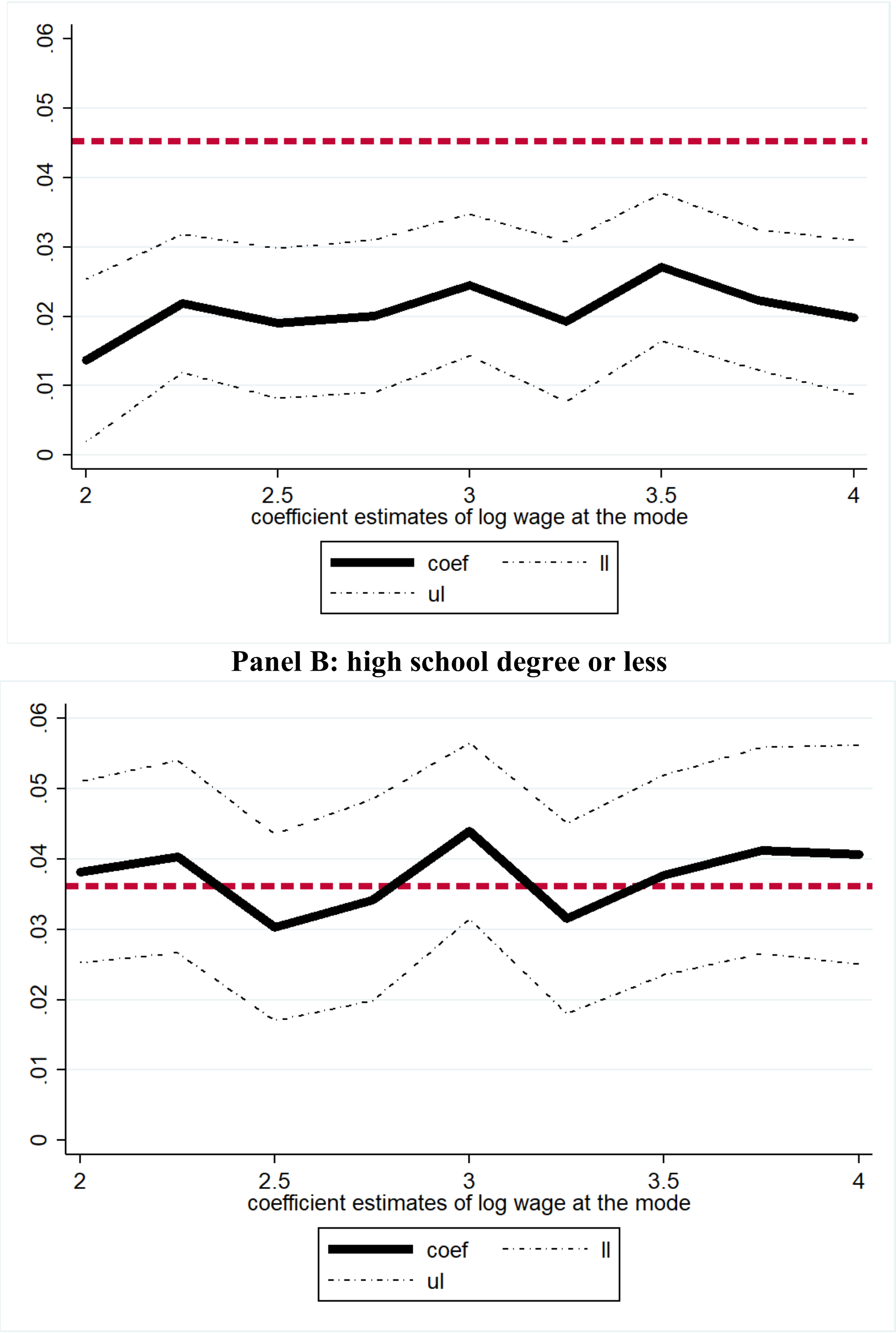


Panel B: high school degree or less



Note: Red dashed horizontal line represents value 0.

Figure 15b: Male modal return robustness check using different bandwidth



Panel A: college degree or more

Panel B: high school degree or less

Note: Red dashed horizontal line represents mean return for each sample.

Chapter 3.

Colonial Legacy, State-building and the Salience of Ethnicity in Sub-Saharan Africa

Merima Ali
Department of Economics
Syracuse University, Syracuse, New York, 13244-1020
maali100@maxwell.syr.edu

Odd-Helge Fjeldstad
CMI - Chr. Michelsen Institute
P.O.Box 6033, N-5892 Bergen, Norway
odd.fjeldstad@cmi.no

Boqian Jiang
Department of Economics and Center for Policy Research
Syracuse University, Syracuse, New York, 13244-1020
bjiang03@syr.edu

Abdulaziz B. Shifa
Department of Economics
Syracuse University, Syracuse, New York, 13244-1020
abshifa@maxwell.syr.edu

Abstract

African colonial history suggests that British colonial rule may have undermined state centralisation due to legacies of ethnic segregation and stronger executive constraints. Using micro-data from anglophone and francophone countries in sub-Saharan Africa, we find that anglophone citizens are less likely to identify themselves in national terms (relative to ethnic terms). To address endogeneity concerns, we utilise regression discontinuity by focusing on observations near anglophone–francophone borders, both across countries and within Cameroon. Evidence on taxation, security and the power of chiefs also suggests weaker state capacity in anglophone countries. These results highlight the legacy of colonial rule on state-building.

JEL Classification: F5, N0, R0

Key words: state capacity, colonial history, Africa.

1. Introduction

Building an effective state that can enforce its laws, maintain stability, and provide public infrastructure remains a major challenge for ethnically diverse countries. Various correlates of ethnic tensions (e.g. fractionalisation, polarisation, segregation, interethnic income inequality) have been shown to be associated with such adverse outcomes as slower economic growth, higher incidence and duration of civil conflicts, weaker state capacity, and the underprovision of public goods.[1] The implications of these empirical patterns are especially severe in Africa owing to its relatively high level of ethnic diversity. Despite the emphasis on the problem of weak state capacity in Africa,[2] the role of history—in particular, that of colonial legacies—is far from fully understood.

In this paper, we focus on the legacy of colonial occupation by the two largest colonial powers—Britain and France—on state building in sub-Saharan Africa. The literature on colonialism and African history suggests two possible reasons why the legacy of British rule may differ from that of French rule. First, Britain adopted a 'divide and rule' strategy in which ethnic identities played a central role. A prominent feature of British colonial rule was its emphasis on *native administration*, a system of decentralised control in which the local population was segregated along tribal lines and ruled indirectly by local chiefs. Native administration empowered chiefs to rule over their respective local populations and instituted a rigid association between one's ethnic identity and access to basic resources (such as land and local government services). In contrast, France's colonial policy featured a more centralised approach in which ethnic cleavages played a less significant role. Local administrative boundaries in French colonies did not necessarily represent specific ethnic groups and, therefore, did not hinder various ethnic groups

[1] See (Mauro(1995), Easterly and Levine(1997), Alesina ~al.(2003)Alesina, Devleeschauwer, Easterly, Kurlat and Wacziarg, Miguel and Gugerty(2005), Montalvo and Reynal-Querol(2005), Montalvo and Reynal-Querol(2005), Baldwin and Huber(2010), Alesina and Zhuravskaya(2011), Hjort(2014), Alesina ~al.(2016)Alesina, Michalopoulos and Papaioannou).

[2] For a detailed review, see (Acemoglu ~al.(2016)Acemoglu, Chaves, Osafo-Kwaako and Robinson).

from belonging to the same political unit. Moreover, the use of French as a common official language was promoted throughout the colonies, encouraging language integration within the colonial bureaucracy. The power of local chiefs was also suppressed. These differences in approach to colonial rule suggest that the legacy of British rule may have undermined the formation of a shared national identity across ethnic groups, empowered local chiefs and weakened the central state.[3]

Second, the French and English legal systems have different implications for the power of the executive. French civil law is often said to leave more political power and control in the hands of the central state, whereas judicial independence from the executive is viewed as a defining feature of British common law (Beck and Levine, 2005; La Porta et al., 2008). Under civil law, the state could have a stronger legal power to centrally organise and control society (through policies such as state ownership of enterprises and military conscription). However, the relatively independent judiciary under common law may work against such a centralisation. Thus, the differences in the executive power under the two legal traditions also suggest that state centralisation might be relatively weaker in anglophone (than in francophone) countries because of colonial 'legal origins'.

Hence, British colonial rule, because of its approach to segregating ethnic groups and the common law tradition, may well have undermined state centralisation. At the same time, it is far from obvious whether this hypothesis is confirmed in the case of Africa (Herbst, 2014). Colonisers faced enormous logistic challenges in the African hinterland owing to tropical diseases and a lack of accessible roads. Herbst argues that the colonisers' effective control over their official territories

[3] See Dilley (1966); Crowder (1968); Miles (1994); Mamdani (1996); Acemoglu et al. (2011); Alesina and Zhuravskaya (2011); Acemoglu et al. (2016).

was quite limited and, as a result, it is far from certain whether differences in colonial policies had a lasting impact on postcolonial state capacity.

In our empirical analyses, the main outcome variable is a micro-level indicator for the strength of national (versus ethnic) identification in a sample of adult Africans. This variable measures the extent to which individuals identify themselves with their countries as opposed to their own ethnic groups. We focus on the strength of national identification as our main outcome variable for two reasons. First, creating a sense of shared identity among citizens—a challenging task in the context of ethnic rivalries—is an important component of state-building (Alesina and Reich, 2013). Second, differences in the approaches to colonial rule and legal origins raise the question of whether colonial legacies affected the sense of national identification in contemporary Africa. One would expect that the legacy of ethnic segregation under British colonial rule fostered interethnic rivalry and hindered the formation of shared national identity among citizens. Furthermore, a stronger executive constraint in the British legal tradition would be expected to limit the state's power to implement national policies aimed at encouraging interethnic integration, such as army conscription, centrally managed elementary education, and an expansive government bureaucracy.

Besides the evidence on national identification, we also examine the relationship between colonial rule and several different proxies for state capacity. The literature on state capacity emphasises the ability to raise taxes and to maintain law and order as important factors in economic development (Besley and Persson, 2009, 2010, 2011). In the African context, it is often argued that powerful traditional chiefs restrict the central state from exercising its control (Acemoglu et al., 2016). Given the prominent role of chiefs under British colonial rule, one would expect chiefs to have greater power in anglophone countries. Relatedly, the existence of more layers of hierarchy prior to contact with Europeans is robustly associated with greater development today (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013, 2014). We therefore examine the empirical evidence on the association between British rule and state-capacity indicators for taxation, security, and the power of traditional chiefs.

We use data from several rounds of the Afrobarometer surveys, which contain information on nationally representative samples of adult citizens in a number of African countries. The surveys provide data on about 100,000 respondents from twelve anglophone countries and nine francophone countries. Preliminary comparisons between all the anglophone and francophone respondents in our data show that anglophone respondents are less likely to identify themselves with their countries than with their ethnic groups, providing suggestive evidence for a negative association between the legacy of British rule and national identification.

As a first step in addressing endogeneity concerns, we exploit the wide geographic coverage of the data set and implement regression discontinuity (RD) analysis focusing on a subset of respondents from areas near the borders between anglophone and francophone countries. Given the arbitrary nature of colonial borders (Michalopoulos and Papaioannou, 2016), the results of this analysis help minimise the concern that precolonial differences between (what ended up being) anglophone and francophone countries could confound the results owing to, for example, a possible correlation between the coloniser's identity and ethnic rivalries that predated colonialism (Besley and Reynal-Querol, 2014). We find that the RD analysis also yields the same result: anglophone respondents report a weaker sense of national identity. Using observations from Cameroon, whose territory was divided between France and Britain, we carry out additional sets of RD analysis in which the variation in colonial status comes from within the same country—

thereby keeping country-level differences constant. The results from the Cameroon sample also show a weaker sense of national identity among anglophone respondents.

The weaker sense of national identity among anglophones is consistent with the two explanations offered previously: a weaker executive power in the British legal tradition and the British approach to colonial rule (with respect to the role of ethnicity in colonial administration). Directly isolating the effects of these two potential explanations is not possible because our independent variable, namely the coloniser's identity, does not distinguish between them. Although the negative association between British rule and the strength of national identification—consistent with the literature on colonialism and African history—remains the main empirical contribution of this paper, our results also provide some evidence suggesting that the approach to colonial rule (as opposed to legal traditions) is the more likely explanation. First, our result from Cameroon relies on within-country variations in the coloniser's identity, suggesting that the difference in legal origins, which varies at country-level, is not the driving reason. Second, we introduce several controls (such as citizens' trust in the judiciary) that are plausibly associated with legal origins. We find that the result is unchanged when these controls are added, suggesting that differences in legal traditions do not affect the results.

We also examine a number of outcome variables to assess the evidence on taxation, chiefs' power, and security. Many of these outcome variables are constructed from Afrobarometer surveys on the experiences and attitudes of respondents regarding taxes, safety, and the role of chiefs in their community. In particular, we use micro-level indicators for tax-compliance norms, the strength of tax enforcement, the prevalence of extortion activities by non-state actors, the prevalence of crimes (such as theft), and the power of traditional chiefs. In addition to the Afrobarometer surveys, we also use two more data sets that provide georeferenced information on conflict events; this information is used to construct indicators for the prevalence of armed conflicts.

The empirical patterns from these outcome variables also suggest weaker state capacity among anglophones. As we show in Section 4, some of the estimated relationships between colonial status and these variables are insignificant while many are significant. For example, of the two security indicators that we sourced from Afrobarometer data, the first (prevalence of crime) is insignificant in some of the specifications whereas the second (prevalence of extortion by non-state actors) is significant across all specifications. Our indicator for the prevalence of armed conflicts is also significantly higher in anglophone regions across most specifications. Although anglophone respondents tend to report weaker tax enforcement and stronger power of chiefs, tax-compliance norms do not show significant differences. Crucially, however, the general pattern is that *all* of the significant coefficients suggest a lower state capacity (i.e. weaker tax enforcement, stronger power of chiefs, and less security) among anglophone countries. Thus, the broad picture portrayed by these results—in line with the literature on colonial legacy and African history—is one that associates the legacy of British rule with weaker state capacity.

This paper contributes to the literature on the role of history in state development, which we complement by highlighting the role of colonial history in state-building.[4] A strand of this literature examines the effect of precolonial history on contemporary development. The seminal study by Nunn (2008) shows how Africa's slave trade is correlated with ethnic fractionalisation, state development, and income (see also Nunn and Puga, 2012). Nunn and Wantchekon (2011) find an adverse effect of slave trade on trust levels; Besley and Reynal-Querol (2014) explore the relationship between precolonial interethnic conflict and the contemporary salience of ethnic

[4] See Nunn (2014) and Michalopoulos and Papaioannou (2015, 2018) for a detailed review of the literature on the role of history in economic development.

identity; and Michalopoulos and Papaioannou (2014) document the impact of precolonial institutions on subnational development.

Another strand of the historical literature studies the role of colonial history in postcolonial development. A number of these studies use macro-level cross-country data (Acemoglu et al., 2001, 2002; La Porta et al., 2008; Feyrer and Sacerdote, 2009). From a methodological standpoint, our paper is closely related to the growing literature that utilise RD analysis of micro-level data to examine colonial legacies (see, e.g., Banerjee and Iyer, 2005; Dell, 2010; Bubb, 2013; Pinkovskiy, 2013; Michalopoulos and Papaioannou, 2014, 2016; Baruah et al., 2017; Lechler and McNamee, 2017).However, none of these studies look at the effect of colonial rule on state-building in Africa.

We also contribute to the literature on the salience of ethnic identity in national politics. Alesina and Reich (2013) develop a theoretical model of state-building in which the construction of national identity (or 'homogenisation of citizens') is endogenised. Citing several historical examples, Leeson (2005) argues that colonial rulers' segregation of local populations along tribal lines disrupted the interethnic cooperation that existed in precolonial Africa, thereby inhibiting the integration of tribes and the formation of shared national identity.[5] Eifert et al. (2010) report that ethnic identity becomes more salient in response to increases in competition for political control (as measured by proximity to election periods). Miguel (2004) compares post-independence outcomes in Kenya and Tanzania to show that government policies, such as public school curriculum and establishing a national language, can promote the formation of stronger national identification. We complement this literature by highlighting the role of colonial history for the salience of ethnic identification. More recently, Blouin and Mukand (2018) find that government propaganda was effective in encouraging greater inter-ethnic trust and cooperation in Rwanda.

[5] See also Leeson (2008) for a formal model incorporating this idea of cooperation through informal institutions.

The rest of our paper is organised as follows. The next section provides a brief historical review of the British and French approaches to colonial rule. Thereafter, results on the strength of national identification are presented in Section 3. In Section 4, we report results on taxation, security, and the power of chiefs. Section 5 concludes.

2. Historical Background

The central component of British colonial rule was ‘native administration’. As outlined by Frederick D. Lugard—the renowned colonial official whose extensive engagement in Africa ranged from being a military commander in Nyasaland (1888) to governing Nigeria (1914–1919)—the main tenet of native administration is that natives should be ‘administered’ by their native chiefs in accordance with their native customs and on their native land (Lugard, 1922). In practice, native administration was a fairly autonomous satellite institution in the hierarchy of colonial bureaucracy, which segregated locals along tribal affiliations and controlled them via locally powerful men (chiefs). Britain adopted this kind of indirect rule as a way of controlling the local population with minimal cost (Chanock, 1985; Okoye, 2017). The more direct rule imposed in Malawi and southern Nigeria during the early periods of colonial occupation (late 19th century) proved to be too costly to sustain when—following the Berlin conference of 1884–1885—the colonial territories expanded vastly to areas where the government had limited control.[6] Hence, native administration was imposed more or less throughout the territories in non-settler colonies, such as Uganda and western Africa. In colonies that contained a large number of settler populations, native administration was applied only to the locals within their reserves (Tinger, 1976; Chanock, 1985).

[6] For a formal model of indirect rule, see Padró I Miquel and Yared (2012).

The first key feature of native administration is the role of chiefs, who wielded significant control over locals. Locally powerful men from the precolonial power structure (such as tribal chiefs) were co-opted into the colonial administration, often by threatened or actual military attacks (Crowder, 1964). In areas where identifiable tribes or tribal chiefs did not exist, the chiefs were 'invented' and imposed on the locals (Khapoya, 2010). Backed by British military support, chiefs ruled the local population and extracted taxes while being held accountable primarily to their colonial master (i.e. the district commissioner). The chief presided over all branches of the local government: he set the rules, acted as a judiciary, and controlled the administration. He also appointed the subchiefs and village headmen.

The second main feature of native administration was segregating the local population along tribal lines, which served to undermine cooperation among various ethnic groups and so reduce the threat of a more unified and stronger resistance against colonial rule (Fanthorpe, 2001). Boundary demarcations of the native administration units assigned collective ownership of land to tribes, with the chief having the ultimate power to decide the allocation of plots among his subjects. Thus, native administration instituted a rigid association between tribal identity and access to basic resources, such as land and local government services.

Many scholars of African history argue that the natives' tribal identity was relatively less prominent in the French colonies (see, e.g., Whittlesey, 1937; Crowder, 1964; Miles, 1994; Mamdani, 1996). First, the areas marked by local administrative boundaries did not necessarily represent specific ethnic groups and often cut across preexisting political boundaries. Thus, they did not prevent various ethnic groups from belonging to the same political unit. This approach was in contrast to British rule, under which colonies were essentially organised as autonomous collections of tribal authorities. Second, although the French also used chiefs in many instances, the power of those chiefs was often suppressed. French colonial law did not give chiefs the power to allocate land among natives. It also retained most of the judicial power with the resident commander (*le commandant de cercle*). Finally, chiefs were allowed relatively less autonomy in appointing subchiefs and village headmen. Devoid of legal power to allocate land, control the local judiciary, and appoint subchiefs, the chief's primary role in French colonies was reduced to executing the commander's orders within an administrative bureaucracy. Hence, the colonial bureaucracy under French rule, as compared with the native administrations in the British colonies, was less dependent on the chief's patronage network.

The prominence of native administration in British colonies meant that political rights were tied to an individual's ethnic identity and not to citizenship, thus undermining the practical relevance of citizenship (Fanthorpe2001). The possibility of excluding others tended to induce competition for resources and political influence along ethnic lines and foster rivalries. Schildkrout (1970, pp. 374–75) notes that, in Ghana, the British demand for Kumasi residents to appoint their own tribal headmen intensified rivalries among the various ethnic groups (e.g. the Hausa, Yoruba, and Mossi) as each group rallied for more influence through its own headman.

Such interethnic rivalries can have a long-term effect on nation building. First, a sense of animosity and mistrust among ethnic groups could persist for an extended period (Nunn and Wantchekon, 2011; Voigtländer and Voth, 2012; Rohner et al., 2013). Furthermore, existing divisions could be exploited by successive generations of politicians through ethnic favouritism, reinforcing the initial cleavages. Finally, the continued prominence of tribal chiefs in postcolonial anglophone countries means that ethnic identity could remain an important political factor.

In addition to undermining the construction of shared national identity, the existence of powerful chiefs may also directly undermine state centralisation. Because the emergence of a strong central state is likely to threaten powerful local chiefs, they could have an incentive to use their power to keep the state weak (Acemoglu et al., 2016).

3. Empirical Results: National Identity

To measure the strength of national identification, we construct a variable based on a survey question about the respondent's sense of national (relative to ethnic) identity. Respondents were asked to describe their sense of identity by choosing one of five options: (1) only ethnic, (2) more ethnic than national, (3) equally ethnic and national, (4) more national than ethnic, or (5) only national. Our outcome variable, *National identity*, is a binary index that equals 1 if the respondent chooses either option (4) or (5)—that is, if the respondent places more importance on national than on ethnic identity. Otherwise, *National identity* equals 0. We obtain similar results when using an alternative index that takes values ranging from 0 to 4, where higher values are assigned to statements corresponding to a greater salience of national identity. The descriptive statistics for *National identity*, and for all other variables to be used in our analysis, are presented in Table 1.

We present our results in three stages, each corresponding to subsamples from different geographic subunits. First, Section 3.1 presents the preliminary results using all observations in our sample, which consists of about 100,000 respondents. These observations are drawn from rounds 3–6 of the Afrobarometer surveys that covered twelve anglophone and nine francophone countries (see Figure 1).

Section 3.2 presents the RD analyses that focus on observations near the anglophone–francophone national borders of western African countries in our sample. Section 3.3 introduces additional controls in order to examine whether the results can be explained by differences in legal origins. Section 3.4 concludes this section by presenting the RD results based on observations from Cameroon.

3.1. Preliminary Results

We consider a model given by the following regression equation:

$$Y_i = \alpha + \beta \times Anglophone_i + \mathbf{X}_i{}'\Gamma + \varepsilon_i,$$

where Y_i is the dependent variable and i indexes the respondent. $Anglophone_i$ is an indicator for whether the respondent is from an anglophone country: it equals 1 for anglophone respondents and 0 for francophones. Our coefficient of interest is β, which captures the difference between anglophone and francophone respondents with respect to the outcome variable. The vector $\mathbf{X}_i$ includes a set of controls; these controls will be described later as they are introduced into the regression estimations.[7] Summary statistics are reported in Table 1, and a detailed description of our data sources for each variable is given in the Appendix.

The first row of Table 1 shows the means and standard deviations of *National identity*. We see that, compared with francophone respondents, the share of anglophone respondents who prioritise national identity is lower by 13 percentage points. That is, 55% of francophone respondents prioritise national identity whereas only 42% of anglophones do so. Table 2 presents the regression results using all observations in our sample. We report robust standard errors clustered at both ethnicity and country levels (Cameron et al., 2011).[8]

The controls include several variables that could affect state-building; therefore, they account for the possibility that the correlation between those variables and colonial status may

[7] Non-linear probability models (logit or probit) yield qualitatively identical results. We report results from our linear model because it is more straightforward in terms of both estimation procedure and interpretation (e.g. estimated effects represent mean percentage differences between francophone and anglophone observations). The linear model is also less sensitive to distributional assumptions concerning the error terms, which is important given our use of several dummy controls (cf. Angrist2008).

[8] The results remain the same when we cluster at district (instead of ethnicity) and country levels.

confound our results. We include controls at national, district, ethnicity, and individual levels. The subnational and individual controls have the advantage of capturing variations across regions and/or individuals. This advantage is especially important in Africa, where state capacity varies significantly across regions because the central states tend to have limited control over areas remotely located from capital cities.

In the first column, we report results with no controls (except fixed effects for survey rounds). Column (2) includes region indicators (eastern, western, and southern Africa) as well as an indicator variable for whether the country is landlocked. To account for the possible effect of German occupation in some African countries, we include an indicator variable for whether a country was a former colony of Germany. One francophone country (Togo) and two anglophone countries (Namibia and mainland Tanzania) were German colonies prior to the First World War;[9] when Germany was defeated, they were transferred to France and Britain.

The third column of Table 2 controls for several individual-level, socio-economic characteristics of the respondents. These variables are all sourced from the Afrobarometer survey. They include: indicator variables for the location of respondents (urban vs. rural) as well as their employment status (employed vs. unemployed) and gender; nine indicators for education levels; controls for age and age squared; eight fixed effects for the respondents' religions; and three indicators for asset ownership.

Column (4) includes the remaining controls. These include a range of variables to account for historical, institutional, demographic, and economic factors. In selecting these controls, we closely follow Nunn and Wantchekon (2011) and Nunn and Puga (2012). To account for

[9] In addition, two anglophone countries—Ghana and Nigeria—had small portions of land transferred from German ownership (British Togo joined Ghana, and part of the British Cameroon joined Nigeria). However, the major portion of these countries was under British rule and so they are not considered to be former German colonies.

precolonial legacies, we control for historical levels of the intensity of exposure to slave trade, urbanisation, and complexity of institutions. The intensity of exposure to slave trade, which eroded trust levels (including interethnic trust), may affect state-building by intensifying conflict and hindering integration across ethnic groups (Nunn and Wantchekon, 2011; Fenske and Kala, 2017). This measure is constructed by dividing the total number of slaves exported from each ethnic group by the size of land area that is historically inhabited by the ethnic group (Nunn and Wantchekon, 2011). Ethnic homelands are defined according to Murdock's (1959) Ethnolinguistic Map. We use village-level geographic data on the residence of each respondent to project the locations of respondents on the Ethhnolinguistic Map.[10] As a further control for exposure to slave trade, we include the distance of each ethnic homeland from the nearest coast. The control for historical levels of urbanisation is an indicator variable for whether the respondent is located in an ethnic homeland that contained, in 1800, a city whose population was at least 20,000 (Chandler and Fox, 1974). This control is meant to account for the extent of precolonial economic development. Since the complexity of precolonial institutions is found to be correlated with contemporary development (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2014), we include four dummies to control for the number of precolonial jurisdictional hierarchies in each ethnic group (Murdock, 1967).

We include two controls to account for colonial activities. The first one, an indicator variable for whether an ethnic homeland had a colonial railway station, is meant to account for colonial investments in infrastructure (Dell and Olken, 2017). Since missionary activities by Europeans—which tended to be more common in British colonies—may have lowered trust levels and weakened traditional institutions (Okoye, 2017), we control for the number of missionary

[10] We use geodata on Afrobarometer respondents from Knutsen et al. (2016) and AidData.

stations per area of each ethnic homeland. Data on these controls are from Nunn and Wantchekon (2011).

In order to account for the potential effect of ethnic composition on interethnic relationships (Easterly and Levine, 1997; Alesina et al., 1999; Alesina and La Fer- rara, 2002), we control for ethnic fractionalisation in the respondent's district and the share of population in the district that is of the same ethnicity as the respondent. In constructing both variables, we follow Nunn and Wantchekon (2011) and use the sample of individuals in the Afrobarometer surveys.

Finally, we include two controls for the level of contemporary economic development in the historical homeland of each ethnic group. Economic development may affect state capacity, as the size of the formal sector tends to increase with the level of economic development (Besley and Persson, 2011). Since reliable income data at subnational levels are not available, we follow the recent literature and use the density of night-time light—based on satellite images—as a proxy for economic activity (Henderson et al., 2012; Michalopoulos and Papaioannou, 2014; Pinkovskiy and Sala-i Martin, 2016). Following Michalopoulos and Papaioannou (2014), we construct a measure of light density per square kilometer for the period 2011–2013 by averaging across pixels that lie within each ethnic group's historical homeland.[11] Since state capacity is likely to be lower in areas farther from the capital city, we control for the distance of each ethnic homeland from the capital.

According to our estimate in column (1) of Table 2, the coefficient for *Anglophone* is negative and statistically significant. The share of respondents who prioritise national identity in anglophone countries is lower by 12 percentage points, which reaffirms the mean difference

[11] This 2011–2013 period corresponds to the years during which round 5 of the Afrobarometer surveys was undertaken.

between anglophone and francophone respondents reported in Table 1. This result remains significant when we incorporate the remaining controls in columns (2)–(4).

3.2. Evidence from Regression Discontinuity

Despite the inclusion of several controls, there may still be endogeneity concerns due to possible confounding factors that are difficult to control. This could, for instance, be the case if Britain's policy of adopting native administration may have led it to target regions that already had a strong sense of ethnic identity. On the other hand, France's lack of such a motive may have led it to focus on regions consisting of relatively homogeneous ethnic groups. If such a selection strategy were operative, then current differences (i.e. between anglophone and francophone countries) in the salience of ethnic identity may have existed prior to colonisation and thus would not reflect colonial legacy.[12] As a first step towards mitigating these endogeneity concerns, we undertake RD analysis on a limited set of respondents who reside near national borders between anglophone and francophone countries. The key assumption required to address the selection problem in the RD analysis is that, prior to colonisation, anglophone and francophone regions within the RD sample were not systematically different in terms of (un)observable factors that could have affected the strength of national identification.

There is little disagreement among scholars of African history that most national borders were drawn arbitrarily by colonisers.[13] The colonisation of Africa happened rapidly. The borders were drawn hastily in European capitals with little knowledge of local situations. As summarised

[12] However, Wesseling (1996, pp. 177–78) argues that the massive French occupation in western Africa was driven more by the navy's desire to redeem itself from past humiliation than by any bona fide strategic concerns.

[13] See Michalopoulos and Papaioannou (2016) for a detailed review of the literature and for evidence on the arbitrariness of African border demarcations by colonial powers.

in the much cited statement by Lord Salisbury (then the British Foreign Secretary, later Prime Minister) at the colonial powers' 1884–1885 'carve-up' conference in Berlin:

> We have been engaged in drawing lines upon maps where no white man's feet have ever trod; we have been giving away mountains and rivers and lakes to each other, only hindered by the small impediment that we never knew exactly where the mountains and rivers and lakes were. (Muiu, 2010, p. 1)

As a result, the borders typically divided communities that belonged to relatively homogeneous groups that shared similar ethnicity, political organisation and agro–economic zones. As illustrated in Figure 2, the arbitrariness of African borders also stands out in our RD sample. This map projects Murdock's (1959) Ethnolinguistic Map on country borders in western Africa that are included in our RD analysis. The thickest lines show the borders between anglophone and francophone countries. Three of the countries along those borders are anglophone (Ghana, Nigeria, and Sierra Leone) and the rest are francophone (Benin, Burkina Faso, Côte d'Ivoire, Guinea, Niger, and Togo). The highlighted portion of the map shows the historical homelands that were split into more than one country along the anglophone–francophone borders. Following Michalopoulos and Papaioannou (2016), we define 'split' groups as historical homelands for which at least 10% of their territories are found on both sides of a national border. A visual inspection of Figure 2 reveals that, with few exceptions, the borders cut through ethnic homelands—affirming that the national borders in our RD sample likewise reflect the largely arbitrary nature of most African borders.[14]

Of the 91 ethnic historical homelands in our RD sample, which covers the areas that lie within 100 kilometers of anglophone–francophone national borders, the majority (51 groups) are split between countries, suggesting that a significant portion of our RD regions come from

[14] If we instead define split groups as ethnic homelands where at least 5% (rather than 10%) of the historical homelands are found on both sides of a national border, then even more portions of the national borders will cut through ethnic homelands (and are thus rendered even more arbitrary).

communities that were unlikely to have systematic precolonial differences across national borders.[15] Comparing the 40 non-split groups in anglophone and francophone countries with respect to various observable precolonial characteristics—such as exposure to slave trade, urbanisation in 1800, and levels of complexity in precolonial political organisations—we find that the differences between those on francophone versus anglophone sides of the borders are statistically insignificant.

Table 3 presents the estimated results. As a benchmark comparison, we begin by reporting a regression result using all francophone and anglophone observations from western Africa; this result is presented in the first column. We report the result with no controls to provide a transparent comparison of the mean difference between all francophone and anglophone respondents in western Africa. However, including the additional controls does not change the result. There are a total of 43,013 observations from western Africa, which account for 44% of all the observations in our data set. Columns (2)–(5) present regression results using only the observations in our RD sample (i.e. observations within 100 km of an anglophone–francophone national border).[16] This RD sample includes nearly 13,000 observations, or about 30% of all observations in western Africa (approximately 13% of the entire sample's observations). Column (2) in the table includes no controls except for survey-round fixed effects and national border fixed effects. Column (3) controls for distance to the anglophone–francophone national borders (on either side). Columns (4) and (5) include, respectively, the individual-level controls and the remaining controls; see Table 2. We discuss column (6) in Section 3.3. All coefficients estimated from these regressions

[15] We consider a historical homeland to be part of our RD sample if its geographic boundary overlaps the RD area.

[16] The results are not sensitive to reasonable alternations of the cut-off (e.g. 60, 80, or 120 km). This insensitivity is intuitive in light of the RD plots, which tend to show that the significant (resp., insignificant) results display (resp., do not display) visible shifts at the border.

are significantly negative, which confirms the previous pattern that the strength of national identification tends to be lower among anglophone respondents.

Figure 3 offers a visual display of the strength of national identification by distance to border. The fitted lines represent the correlation between distance to national borders and *National identity* along with their 95% confidence intervals (from an OLS regression of *National identity* on distance). The dots mark local averages (in 10-km bins) of *National identity* and represent the share of respondents who identify more with their country than with their ethnic group. The advantage of an RD plot is that it provides a more transparent characterisation of the data. These patterns are consistent with the findings reported in Table 3 on the anglophone side of the borders (to the right of the x-axis center point), the level of national identification tends to be lower.

3.3. Accounting for Legal Origin

As discussed in the Introduction, the literature on colonial legacy and African history suggests two possible factors—namely, differences in approaches to colonial rule and legal origin—that may associate the coloniser's identity with state building. Directly isolating the effects of these factors is not feasible because our independent variable (i.e. the identity of the coloniser) does not distinguish between the two effects. Nevertheless, we examine this question by controlling for variables that are reasonably presumed to be associated with legal origins. If the results remain the same when we control for these variables, this provides suggestive evidence that the effects are less likely to be driven by differences in legal origin.

The literature on colonial legacies emphasises that the British legal origin tends to be associated with greater judicial independence than does the French (La Porta et al., 2004). This judicial constraint on the executive's power could limit the latter's ability to control society and thus to strengthen the central state. The transmission of both institutions' features—including legal codes and such cultural values as attitudes towards political freedom—from colonisers to colonies have been posited as channels through which legal origin may affect judicial independence (La Porta et al., 2008). We therefore use three variables from the Afrobarometer data to control for potential differences concerning the judiciary and attitudes towards political freedom.

The first variable measures the level of respondents' trust in 'the courts of law'. Respondents were asked to choose one of the following options to express their level of trust in the courts: (1) not at all, (2) just a little, (3) somewhat, and (4) a lot. Using this variable, we construct a binary control that indicates whether the respondent chose one of the last two options or rather one of the first two. We also checked the robustness of our results to alternative ways of defining the controls (e.g. fixed effects for each type of response); we obtain similar results.

The other two variables measure respondents' attitudes about political freedom. One of them is based on a survey question that asked respondents to give their view of these two statements: (A) 'We should be able to join any organization, whether or not the government approves of it'; and (B) 'Government should be able to ban any organization that goes against its policies'. The other variable is constructed from respondents' answers to questions probing their views about press freedom. They were asked to describe their attitude towards the following two statements: (A) 'The media should have the right to publish any views and ideas without government control'; and (B) 'The government should have the right to prevent the media from publishing things that it considers harmful to society'. To each of these two questions, the respondents answered by choosing one of five options: (1) agree very strongly with Statement A, (2) agree with Statement A, (3) agree with Statement B, (4) agree very strongly with Statement B, or (5) agree with neither statement. Our indicators for respondents' attitudes about freedom include two binary variables, one for each question, indicating whether they chose options (1) or (2)—that is, agreeing with statements favouring greater freedom—or rather options (3), (4), or (5).

Column (6) of Table 3 reports estimation results (for the RD sample) when we include the controls for trust in courts of law and attitudes towards political freedom. The coefficient for *Anglophone* remains essentially the same, providing no indication that differences in legal origins—as measured by trust in courts and attitudes towards political freedom—are driving the results.

3.4. Evidence from Cameroon

Cameroon was first colonised by Germany in the mid-1880s. Following Germany's defeat in the First World War, Britain and France each controlled portions of Cameroon and split it into two parts in 1919. Sections of south-western and northwestern Cameroon (bordering Nigeria) became part of the British colony while the rest was colonised by France. After independence, the two parts of Cameroon (except for the north-western part that joined Nigeria) reunited in 1961 to form Cameroon as it is currently configured. After this reunification, Cameroon endured the strongly authoritarian rule of President Ahmadu Ahidjo, whose Cameroon National Union was the sole legal party during much of his rule. The Cameroon state's authoritarian nature has essentially continued until today under President Paul Biya, who succeeded Ahidjo in 1982. Freedom House has classified Cameroon as 'Not Free' ever since 1999, the first year for which data were available.[17]

[17] Polity IV has likewise assigned negative scores in each of the years since 1960. Both Papaioannou and Siourounis (2008) and Acemoglu et al. (2017) categorize Cameroon as a non- democracy.

The Cameroon case offers a useful setting for RD analysis. First, the variation in colonial rule comes from within Cameroon, allowing us to hold country-level differences constant.[18] Second, the colonial borders separating anglophone and francophone parts appear to be quite arbitrary. Like most colonial borders in Africa, Cameroon was partitioned based on hastily arranged agreements.[19] Emphasising this arbitrariness in the demarcations, Lee and Schultz (2012, p. 372) observe that 'the most notable feature of the colonial border was the degree to which it cut across existing ethnic and religious boundaries'. This feature is evident also in Figure 4, where the map of Cameroon is projected onto Murdock's (1959) Ethnolinguistic Map. The thick broken line (within the outlined western territory) represents the anglophone–francophone border in Cameroon, and the highlighted regions represent historical homelands of ethnic groups that were split between francophone and anglophone Cameroon. We see that almost the entire anglophone–francophone border cuts through historical homelands of ethnic groups. Third, there is a broad consensus in the historical literature that the distinction between the French and British approaches to colonial rule in Cameroon is very similar to the broader pattern observed in western Africa (see Section 2). In effect, the British ruled Cameroon as an extension of Nigeria and instituted native administration. Comparing the British and French colonial rule in Cameroon, Chiabi (1997, p. 27) notes that the 'British had to determine who chiefs were and which areas constituted their jurisdictions. This . . . occupied the British for a greater part of the interwar period. Meanwhile, the French took a different approach. Assessment reports to restructure the country along the lines of chiefdoms was not necessary.' As in most of their colonies in western Africa, the French undermined the chiefs' autonomy, 'treating them as petty bureaucrats who can be hired and fired at will' (Lee and Schultz, 2012, p. 375). The legacy of this distinction was noticed soon after independence, when

[18] According to La Porta et al. (1999), Cameroon has a civil law legal tradition.

[19] The border was established in March 1916 (Elango, 2014).

anglophone Cameroon 'maintained their House of Chiefs. Their counterparts in the French tradition saw no need for a House of Chiefs and abolished it in 1971' (Chiabi, 1997, p. 22).

The data are drawn from the last two rounds (i.e. rounds 5 and 6) of the Afrobarometer surveys, which included Cameroon. The country has ten administrative regions; of these, four are contiguous with the anglophone–francophone border. In Figure 4, the western part of the focal area includes these four regions, and the observations from those regions constitute the sample used for our RD analysis.[20]

The RD estimates using the Cameroon data are reported in Table ??. Column (1) presents a benchmark comparison from the sample consisting of all respondents in Cameroon, and columns (2)–(6) present results from the RD sample. Column (2) reports the estimated results in which we include no controls; the controls listed in the table are progressively added in the subsequent columns. Results from the Cameroon sample are similar to our previous results: the sense of national identity is significantly lower among anglophone respondents. The RD plot for the Cameroon sample is displayed in Figure 5, where again the dots mark local averages (in 10-km bins) of *National identity*. This RD plot, too, indicates that the strength of national identification is lower among anglophone respondents.

4. Additional Results: Taxation, Security, and the Power of Chiefs

This section presents the empirical evidence on indicators of state capacity that are related to taxation, security, and the power of chiefs. We first present results for outcome variables that are sourced from the Afrobarometer data set. These variables provide information on the experiences and attitudes of respondents regarding taxation, safety, and the role of chiefs. We then

[20] We include all observations within those regions because Cameroon's anglophone part is quite small; the maximum distance from the anglophone–francophone border (before it crosses into Nigeria) is less than 100 km.

present results for security indicators constructed by using additional data sets that provide geocoded data on incidents of armed conflict. The descriptive statistics for these outcome variables are presented in Table 1.

For all outcome variables, we report results for the whole sample in addition to RD results for observations near the anglophone–francophone national borders in western Africa. Since conflicts tend to be rare events, conflict data do not provide sufficient variation to undertake RD analysis within Cameroon.[21] Hence, RD results from Cameroon are limited to the outcome variables that we source from the Afrobarometer data set.

4.1. Results from Afrobarometer Data

4.1.1. Outcome variables

Taxation. We consider two outcome variables on taxation, *Compliance norm* and *Evasion difficulty*. The data used to construct each variable are sourced from rounds 5 and 6 of the Afrobarometer surveys. *Compliance norm* measures respondents' moral views about tax evasion; it is therefore viewed as being indicative of the strength of social norms against tax evasion. Respondents describe their views regarding tax evasion by choosing one of three ranked statements: tax evasion is (1) not wrong at all, (2) wrong but understandable, or (3) wrong and punishable. We construct *Compliance norm* as a binary index, which is set equal to 1 if the respondent chooses statement (3) and is set to 0 otherwise.[22]

Evasion difficulty is meant to measure the state's ability, as perceived by survey respondents, to enforce tax compliance. Based on their experiences with government services,

[21] Except for some positive values for *One-sided violence* from UCDP, both ACLED and UCDP report zero conflict incidents in Cameroon.

[22] Most of the outcome variables are binary indicators. However, alternative indexes that allow for more than two values—and with respect to which we apply ordered logit or probit models—yield qualitatively similar results.

respondents were asked to describe evading taxes as being (1) very easy, (2) easy, (3) difficult, or (4) very difficult. We set *Evasion difficulty* to 1 if the respondent chooses response (3) or (4) and set it to 0 otherwise.

Security. We consider two indicators as proxies for security: *Extortion prevalence* and *Crime prevalence*. The former is an indicator for the extent to which the state protects its citizens from extortion by non-state actors. Respondents were asked 'how often powerful groups other than the government, such as criminals or gangs, forced people in their community to pay them in return for protecting them, their property or their businesses'. Respondents answered by choosing one of four options: (1) never, (2) only once, (3) a few times, or (4) often. We set *Extortion prevalence* to 0 if the respondent replied 'never' but to 1 for all other responses. Data on this variable are from round 5 of the survey.

Crime prevalence is constructed using respondents' descriptions of their experience (over the twelve months preceding the survey date) with the following three incidents: (1) they feared crime in their home, (2) something was stolen from their house, and (3) they were attacked. *Crime prevalence* assumes values ranging from 0 to 3, corresponding to the number of incidents regarding which respondents answered in the affirmative. *Crime prevalence* equals 3 if the respondent affirms having experienced all three incidents, equals 2 if two of them are reported and so forth. Data on this variable are from rounds 4–6 of the survey.

Power of chiefs. Respondents were asked how frequently they had contacted local chiefs during the twelve months preceding the survey date. As a measure of chiefs' power, we construct the binary index *Chief contact*, which is set equal to 1 if the respondent reports contacting a chief at least once; otherwise, the variable is set equal to 0. Data for this indicator are from rounds 3, 4 and 6 of the Afrobarometer survey.

4.1.2. Results

Table 5 presents the estimates using all observations in our sample. The order in which controls are added does not affect the results; therefore, we report only the results with the full set of controls. The coefficient for *Compliance norm* is not significant. All the other coefficients are significant, and they indicate a weaker state capacity among anglophones. Compared to francophone respondents, anglophone citizens are more likely to report that tax evasion is easier, extortion activities are prevalent, they experience crime incidents, and they contact chiefs.

Table 6 presents estimation results using observations from western Africa. In Panel A, we include all of these observations; in Panel B, we use only those observations from the RD region (i.e. observations from areas within 100 km of the anglophone–francophone national borders). The coefficients for *Compliance norm* and *Crime prevalence* are insignificant in both panels. The remaining coefficients—for *Evasion difficulty*, *Extortion prevalence*, and *Chief contact*—are all significant. So except for *Crime prevalence*, the results reported in both panels are in line with our previous findings estimated for the entire data set (see Table 5)

The corresponding RD plots are presented in Figure B 1 (see Appendix). *Compliance norm* and *Crime prevalence* do not appear to have a significant discontinuity at the border, reaffirming the insignificant coefficients in Table 6, For the other variables, the directions of discontinuities in the RD plots are also consistent with the reported coefficients.

Table 7 presents the RD results from the Cameroon data—that is, the RD observations from the four administrative regions bordering the francophone–anglophone border in Cameroon (see Section 3.4). The coefficients for *Compliance norm* and *Extortion prevalence* retain their earlier patterns—the (former) latter continues to be (in)significant. The coefficients for *Evasion difficulty* and *Chief contact* now lose significance while the coefficient for *Crime prevalence* becomes significant. The RD plots are presented in Figure B 2 (see Appendix). The discontinuity patterns displayed in that figure appear to be consistent with the coefficients estimated in Table 7. Except for *Extortion prevalence* and *Crime prevalence*, which are found to be significant in Table 7 the other variables do not show discernible discontinuities at the border.

Overall, these findings suggest weaker state capacity among anglophones. Given the relatively large number of outcome variables we examine, it is unsurprising that some of the estimated relationships between colonial status and these outcome variables are insignificant. Moreover, our indicators are bound to be imperfect owing to the inherent difficulty of measuring state capacity. The more important pattern, however, is that *all* of the significant coefficients suggest a lower state capacity among anglophones (i.e. weaker tax enforcement, stronger power of chiefs, and less security).

4.2. Results from Conflict Data

4.2.1 Outcome Variable

The incapacity of weak states to contain armed conflicts often poses a significant security challenge. Therefore, we examine the prevalence of armed conflicts as an additional outcome. We use two data sources to construct conflict indicators. The first one is the Uppsala Conflict Data Program (UCDP) Georeferenced Events Dataset Version 17.1, which reports conflict events along with information about the date and geolocation (latitude and longitude) of the events (Sundberg and Melander, 2013; Croicu and Sundberg, 2017). The most recent version of this data set covers the period 1989–2016. Using the information on the longitude and latitude of each conflict event, we aggregate the conflict data into 0.25×0.25 degree cells (approximately 28 km ^{2}) and construct conflict indicators at the grid-cell level.[23] There are 12,843 such grid cells in all of the francophone and anglophone countries in our sample (5,513 in the former and 7,330 in the latter). We then construct indicators for incidence of conflict events in each grid cell. Our conflict variable (defined at the grid-cell level) is a dummy indicator set equal to 1 if a conflict event occurred in the grid cell during the 1989–2016 sample period (and set to 0 otherwise).[24]

The data set also provides information on characteristics of the actors on both sides of the conflict, such as whether the conflict was between non-state actors (e.g. rebel militias) or whether it involved the state. We use this information to construct three conflict variables that vary by type of conflict actors: *State violence*, *Non-state violence*, and *One-sided violence*. *One-sided violence* is a dummy variable set equal to 1 if there was a conflict event in which an armed group attacked unarmed civilians. *State violence* and *Non-state violence* represent conflict events in which both sides were armed. We set *State violence* to 1 if there was a conflict event involving the state (and to 0 otherwise); *Non-state violence* is set to 1 if there was a conflict in which the actors on both sides were non-state groups.

As a robustness check, we also construct these three indicators using the data provided by the Armed Conflict Location Events Dataset (ACLED) Version 7 (Raleigh et al., 2010; Raleigh and Dowd, 2017). One important difference between the two data sets is that UCDP includes conflict incidents that result in at least one fatality whereas ACLED does not exclude nonfatal events (such as injuries) from its domain. Hence, the number of conflict events per period tends to

[23] There is no universally accepted rule on the choice of grid-cell dimensions. For example, (Berman ~al.(2017)Berman, Couttenier, Rohner and Thoenig), in their study of conflict using the same data sets, aggregate into (0.5 × 0.5)-degree cells. We thus checked robustness of our results by aggregating the data into 0.5×0.5 degree cells. We report the results using 0.25×0.25 degree cells because they deliver a more accurate representation of distance from the borders for our RD analysis.

[24] As a robustness check, instead of a dummy for whether any conflict occurred, we used the alternative outcome variables of (a) the number of conflict events in each cell and (b) the number of years with at least one conflict event. Our results remain the same.

be larger in ACLED owing to its wider coverage. The most recent version of ACLED covers the period 1997–2016.[25]

4.2.2 Results

Table 8 reports regression results on conflict outcomes. Since all of our controls except the individual-level ones are location-level variables (e.g. ethnicity-level controls), we generate the same set of controls for the conflict regressions by assigning to each grid cell the values of the geographic unit to which the grid cell belongs.[26] However, distance controls (i.e. distances to the capital city, the nearest coast, and the border) are reconstructed at the grid level. All of these controls are included in the regressions. The results are not sensitive to omitting the controls (or subsets of them). Standard errors are clustered at both ethnicity and country levels.

In Panel A of the table we report the results using all observations; that is, all of the grid cells in anglophone and francophone countries are included. Panel B includes only the countries from western Africa, which account for nearly half of the total observations. Finally, Panel C presents the RD results, for which the sample observations include only those grid cells whose centroids lie within 100 km of the anglophone–francophone national borders in western Africa. Figure B 3 presents the RD plots (see Appendix).

The results from all subsamples show that conflict events are more likely to occur in anglophone regions. The discontinuities revealed by the RD plots are in line with the estimated coefficients. Thus, the results on armed conflicts also suggest weaker state capacity among anglophones.

[25] See Eck (2012) for a detailed comparison of the two data sets. Whereas UCDP explicitly categorises events into the three categories (state, non-state, and one-sided violence), ACLED does not directly classify events in that manner. Instead, it provides data on the types of actors involved, which we use to distinguish between state and non-state violence.

[26] A grid cell belongs to a geographic unit (e.g. ethnic homeland or country) if the cell's centroid lies within the unit's geographic boundary.

5. Conclusion

Building an effective state remains a major challenge for many developing countries. The literature on colonialism and African history suggests two main reasons why the legacy of British colonial rule (as compared with French rule) may contribute to weak state capacity. First, Britain adopted a system of decentralised rule that empowered chiefs over the local population and instituted a rigid association between an individual's ethnic identity and access to basic resources (e.g. land and local government services). Neither the salience of ethnic identity nor the power of traditional chiefs were as crucial under French colonial rule. Second, the French legal system is argued to concentrate more political power in the hands of the central state.

Consistent with this hypothesis, we find a negative relationship between British rule and the strength of national identification. Citizens of anglophone (as compared with francophone) countries report a weaker sense of national identity than of ethnic identity. This finding holds in the sample of all observations in our data as well as from the RD analysis focusing on observations near the anglophone–francophone borders, both across countries and within Cameroon.

We also explore the empirical patterns for various indicators related to taxation, security, and the power of traditional chiefs. All the significant coefficients on these outcomes indicate lower state capacity among anglophones. Thus, the broad pattern from these results is also one that associates the legacy of British rule with weaker state capacity.

This evidence highlights the legacy of colonial rule on state building.

References

Acemoglu, D., Cantoni, D., Johnson, S. and Robinson, J.A. (2011). 'The consequences of radical reform: the French Revolution', *American Economic Review*, vol. 101(7), pp. 3286–307.

Acemoglu, D., Chaves, I.N., Osafo-Kwaako, P. and Robinson, J.A. (2016). 'Indirect rule and state weakness in Africa: Sierra Leone in comparative perspective', in (S. Edwards, S. Johnson and D. N. Weil, eds.), African Successes: Sustainable Growth, pp. 343–370, vol. IV of *NBER Chapters, Chicago*: University of Chicago Press by the National Bureau of Economic Research (NBER).

Acemoglu, D., Johnson, S. and Robinson, J.A. (2001). 'The colonial origins of comparative development: an empirical investigation', *American Economic Review*, vol. 91(5), pp. 1369–401.

Acemoglu, D., Johnson, S. and Robinson, J.A. (2002). 'Reversal of fortune: geography and institutions in the making of the modern world income distribution', *Quarterly Journal of Economics*, vol. 117(4), pp. 1231–94.

Acemoglu, D., Naidu, S., Restrepo, P. and Robinson, J. (2017). 'Democracy does cause growth', *Journal of Political Economy*, vol. forthcoming.

Alesina, A., Baqir, R. and Easterly, W. (1999). 'Public goods and ethnic divisions',*The Quarterly Journal of Economics*, vol. 114(4), pp. 1243–84.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and Wacziarg, R. (2003). 'Fractionalization', *Journal of Economic Growth*, vol. 8(2), pp. 155–94.

Alesina, A. and La Ferrara, E. (2002). 'Who trusts others?', *Journal of Public Economics,* vol. 85(2), pp. 207–34.

Alesina, A., Michalopoulos, S. and Papaioannou, E. (2016). 'Ethnic inequality', *Journal of Political Economy*, vol. 124(2), pp. 428–88.

Alesina, A. and Reich, B. (2013). 'Nation building', *National Bureau of Economic Research Working Paper Series*, (18839).

Alesina, A. and Zhuravskaya, E. (2011). 'Segregation and the quality of government in a cross section of countries', *American Economic Review*, vol. 101(5), pp. 1872–911.

Angrist, J. and Pischke, J. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Baldwin, K. and Huber, J.D. (2010). 'Economic versus cultural differences: forms of ethnic diversity and public goods provision', *American Political Science Review,* vol. 104, pp. 644–62.

Banerjee, A. and Iyer, L. (2005). ‘History, institutions, and economic performance: the legacy of colonial land tenure systems in India’, *American Economic Review,* vol. 95(4), pp. 1190–213.

Baruah, N.G., Henderson, J.V. and Peng, C. (2017). ‘Colonial legacies: Shaping African cities’, *SERC/Urban and Spatial Programme Discussion Paper.*

Beck, T. and Levine, R. (2005). *Legal Institutions and Financial Development*, Boston, MA: Springer US, pp. 251–78.

Berman, N., Couttenier, M., Rohner, D. and Thoenig, M. (2017). ‘This mine is mine! how minerals fuel conflicts in Africa’, *American Economic Review*, vol. 107(6), pp. 1564–610.

Besley, T. and Persson, T. (2009). ‘The origins of state capacity: property rights, taxation, and politics’, *American Economic Review*, vol. 99(4), pp. 1218–44.

Besley, T. and Persson, T. (2010). ‘State capacity, conflict, and development’, *Econometrica*, vol. 78(1), pp. 1–34.

Besley, T. and Persson, T. (2011). *Pillars of Prosperity: the Political Economics of Development Clusters*, Princeton University Press.

Besley, T. and Reynal-Querol, M. (2014). ‘The legacy of historical conflict: evidence from Africa’, *American Political Science Review*, vol. 108, pp. 319–36.

Blouin, A. and Mukand, S.W. (2018). ‘Erasing ethnicity? Propaganda and nation building in Rwanda’, *Journal of Political Economy*, forthcoming.

Bratton, M. and van de Walle, N. (1997). *Democratic Experiments in Africa: Regime Transitions in Comparative Perspective,* Cambridge Studies in Comparative Politics, Cambridge University Press.

Bubb, R. (2013). ‘The evolution of property rights: state law or informal norms?’, *Journal of Law and Economics*, vol. 56(3), pp. pp. 555–94.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. (2011). ‘Robust inference with multiway clustering’, *Journal of Business & Economic Statistics*, vol. 29(2), pp. 238–49.

Chandler, T. and Fox, G. (1974). 3000 *Years of Urban Growth, Studies in Population,* Academic Press.

Chanock, M. (1985). Law, Custom and Social Order: *The Colonial Experience in Malawi and Zambia,* Cambridge University Press.

Chiabi, E. (1997). *The Making of Modern Cameroon*, vol. 1, Lanham, MD: University Press of America.

Croicu, M. and Sundberg, R. (2017). UCDP GED Codebook version 17.1, Department of Peace and Conflict Research, Uppsala University.

Crowder, M. (1964). ‘Indirect rule: French and British style’, *Africa: Journal of the International African Institute*, vol. 34(3), pp. 197–205.

Crowder, M. (1968). West Africa under Colonial Rule, London: Hutchinson. Dell, M. (2010). ‘The persistent effects of Peru’s mining mita’, *Econometrica*, vol. 78(6), pp. 1863–903.
Dell, M. and Olken, B. (2017). ‘*The development effects of the extractive colonial economy: The Dutch cultivation system in Java’*, memo.

Dilley, M.R. (1966). British Policy in Kenya Colony, London, Cass.

Easterly, W. and Levine, R. (1997). ‘Africa’s growth tragedy: policies and ethnic divisions’, *Quarterly Journal of Economics*, vol. 112(4), pp. 1203–250.

Eck, K. (2012). ‘In data we trust? a comparison of UCDP GED and ACLED conflict events datasets’, *Cooperation and Conflict*, vol. 47(1), pp. 124–41.

Eifert, B., Miguel, E. and Posner, D.N. (2010). ‘Political competition and ethnic identification in Africa’, *American Journal of Political Science*, vol. 54(2), pp. 494–510.

Elango, L.Z. (2014). ‘Anglo-French negotiations concerning Cameroon during World War I, 1914-1916: occupation, “condominium” and partition’, *Journal of Global Initiatives: Policy, Pedagogy, Perspective,* vol. 9(2).

Fanthorpe, R. (2001). ‘Neither citizen nor subject? “Lumpen” agency and the legacy of native administration in Sierra Leone’, *African Affairs*, vol. 100(400), pp. 363–86.

Fenske, J. and Kala, N. (2017). ‘1807: Economic shocks, conflict and the slave trade’, *Journal of Development Economics*, vol. 126(Supplement C), pp. 66 – 76.

Feyrer, J. and Sacerdote, B. (2009). ‘Colonialism and modern income: islands as natural experiments’, *The Review of Economics and Statistics*, vol. 91(2), pp. 245–62.

Gennaioli, N. and Rainer, I. (2007). ‘The modern impact of precolonial centralization in Africa’, *Journal of Economic Growth*, vol. 12(3), pp. 185–234.

Henderson, J.V., Storeygard, A. and Weil, D.N. (2012). ‘Measuring economic growth from outer space’, *American Economic Review*, vol. 102(2), pp. 994– 1028.

Herbst, J. (2014). States and Power in Africa: Comparative Lessons in Authority and Control, *Princeton Studies in International History and Politics*, Princeton University Press.

Hjort, J. (2014). ‘Ethnic divisions and production in firms’, *The Quarterly Journal of Economics*, vol. 129(4), pp. 1899–946.

Khapoya, V. (2010). *The African Experience: An Introduction*, Longman.

Knutsen, C.H., Kotsadam, A., Olsen, E.H. and Wig, T. (2016). 'Mining and local corruption in Africa', *American Journal of Political Science*, vol. 61(2), pp. 320–34.

La Porta, R., de Silanes, F.L., Pop-Eleches, C. and Shleifer, A. (2004). 'Judicial checks and balances', *Journal of Political Economy*, vol. 112(2), pp. 445–70.

La Porta, R., Lopez-de Silanes, F. and Shleifer, A. (2008). 'The economic consequences of legal origins', *Journal of Economic Literature*, vol. 46(2), pp. 285–332.

La Porta, R., Lopez-de Silanes, F., Shleifer, A. and Vishny, R. (1999). 'The quality of government', *Journal of Law, Economics, and Organization*, vol. 15(1), pp. 222–79.

Lechler, M. and McNamee, L. (2017). 'Decentralized despotism? indirect colonial rule undermines contemporary democratic attitudes', Munich Discussion Paper No. 2017-7.

Lee, A. and Schultz, K.A. (2012). 'Comparing British and French colonial legacies: a discontinuity analysis of Cameroon', *Quarterly Journal of Political Science*, vol. 7(4), pp. 365–410.

Leeson, P.T. (2005). 'Endogenizing fractionalization', *Journal of Institutional Economics,* vol. 1(1), pp. 75–98.

Leeson, P.T. (2008). 'Social distance and self-enforcing exchange', *Journal of Legal Studies*, vol. 37(1), pp. 161–88.

Lugard, F.D. (1922). *The Dual Mandate in British Tropical Africa*, London: W. Blackwood and Sons.

Mamdani, M. (1996). Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism, Princeton, New Jersey: Princeton University Press.

Mauro, P. (1995). 'Corruption and growth', *Quarterly Journal of Economics*, vol.110(3), pp. 681–712.

Michalopoulos, S. and Papaioannou, E. (2013). 'Pre-colonial ethnic institutions and contemporary African development', *Econometrica*, vol. 81(1), pp. 113–52.

Michalopoulos, S. and Papaioannou, E. (2014). 'National institutions and subnational development in Africa', *Quarterly Journal of Economics*, vol. 129(1), pp. 151–213.

Michalopoulos, S. and Papaioannou, E. (2015). 'On the ethnic origins of African development: traditional chiefs and pre-colonial political centralization', *Academy of Management Perspective*, vol. 29(132–71).

Michalopoulos, S. and Papaioannou, E. (2016). 'The long-run effects of the scramble for Africa', *American Economic Review*, vol. 106(7), pp. 1802–48.

Michalopoulos, S. and Papaioannou, E. (2018). 'Historical legacies and African development', *Journal of Economic Literature*, forthcoming.

Miguel, E. (2004). 'Tribe or nation? Nation building and public goods in Kenya versus Tanzania', *World Politics*, vol. 56(3), pp. pp. 327–362.

Miguel, E. and Gugerty, M.K. (2005). 'Ethnic diversity, social sanctions, and public goods in Kenya', *Journal of Public Economics*, vol. 89(11-12), pp. 2325–68.

Miles, W.F. (1994). *Hausaland Divided: Colonialism and Independence in Nigeria and Niger,* Ithaca and London: Cornell University Press.

Montalvo, J.G. and Reynal-Querol, M. (2005a). 'Ethnic diversity and economic development', *Journal of Development Economics*, vol. 76(2), pp. 293–323.

Montalvo, J.G. and Reynal-Querol, M. (2005b). 'Ethnic polarization, potential conflict, and civil wars', *American Economic Review*, vol. 95(3), pp. 796–816.

Muiu, M.w. (2010). 'Colonial and postcolonial state and development in Africa', *Social Research*, vol. 77(4), pp. 1311–38.

Murdock, G. (1959). *Africa: Its Peoples and their Culture History*, McGraw-Hill. Murdock, G. (1967). Ethnographic Atlas, Pittsburgh, PA: University of Pittsburgh Press.

Nunn, N. (2008). 'The long-term effects of Africa's slave trades', *Quarterly Journal of Economics*, vol. 123(1), pp. 139–76.

Nunn, N. (2010). 'Religious conversion in colonial Africa', *American Economic Review Papers and Proceedings,* vol. 100(2), pp. 147–52.

Nunn, N. (2014). Historical Development, *Handbook of Economic Growth*, vol. 2, North-Holland, pp. 347–402.

Nunn, N. and Puga, D. (2012). 'Ruggedness: the blessing of bad geography in Africa', *Review of Economics and Statistics*, vol. 94(1), pp. 20–36.

Nunn, N. and Wantchekon, L. (2011). 'The slave trade and the origins of mistrust in Africa', *American Economic Review,* vol. 101(7), pp. 3221–52.

Okoye, D. (2017). '*Things fall apart? missions, institutions, and interpersonal trust'*, memo.

Padró I Miquel, G. and Yared, P. (2012). 'The political economy of indirect control', *The Quarterly Journal of Economics*, vol. 127(2), pp. 947–1015.

Papaioannou, E. and Siourounis, G. (2008). 'Democratisation and growth', *The Economic Journal,* vol. 118(532), pp. 1520–51.

Pinkovskiy, M. and Sala-i Martin, X. (2016). 'Lights, camera....income! Illuminating the national accounts-household surveys debate', *Quarterly Journal of Economics*.

Pinkovskiy, M.L. (2013). '*Economic discontinuities at borders: evidence from satellite data on lights at night*', . Raleigh, C. and Dowd, C. (2017). Armed Conflict Location and Event Data Project (ACLED) Codebook.

Raleigh, C., Linke, A., Hegre, H. and Karlsen, J. (2010). 'Introducing ACLED: An armed conflict location and event dataset', *Journal of Peace Research*, vol. 47(5), pp. 651–60.

Rohner, D., Thoenig, M. and Zilibotti, F. (2013). 'War signals: A theory of trade, trust, and conflict', *The Review of Economic Studies*, vol. 80(3 (284)), pp. 1114– 1147.

Schildkrout, E. (1970). 'Government and chiefs in Kumasi Zongo', in (M. Crowder and O. Ikime, eds.), *West African Chiefs – Their Changing Status under Colonial Rule and Independence*, New York: Africana Pub. Corp.

Sundberg, R. and Melander, E. (2013). 'Introducing the UCDP georeferenced event dataset', *Journal of Peace Research*, vol. 50(4), pp. 523–32.

Tinger, R.L.L. (1976). *Colonial Transformation of Kenya: The Kamba, Kikuyu, and Maasai from 1900 to 1939*, Princeton, NJ:: Princeton University Press.

Voigtla¨nder, N. and Voth, H.J. (2012). 'Persecution perpetuated: the medieval origins of anti-semitic violence in Nazi Germany', *Quarterly Journal of Economics*, vol. 127(3), pp. 1339–92.

Wesseling, H. (1996). *Divide and Rule: The Partition of Africa, 1880-1914*, Praeger

Whittlesey, D. (1937). 'British and French colonial technique in West Africa', *Foreign Affairs*, vol. 15(2), pp. 362–73.

Table 1: Descriptive Statistics

	Obs.		Mean		Survey Rounds
	Anglo.	Franco.	Anglo.	Franco.	
	(1)	(2)	(3)	(4)	(5)
Main outcome variable					
National identity	68,807	30,036	0.42 (0.49)	0.55 (0.50)	3,4,5,6
Outcome variables on taxation, security and chiefs' power					
Outcome variables from Afrobarometer surveys					
Compliance norm	39,270	19,433	0.51 (0.50)	0.58 (0.49)	5,6
Evasion difficulty	35,970	18,131	0.79 (0.41)	0.80 (0.40)	5,6
Extortion prevalence	19,097	8,948	0.13 (0.33)	0.04 (0.20)	5
Crime prevalence	55,564	25,390	0.78 (0.94)	0.57 (0.78)	4,5,6
Chief contact	49,067	20,783	0.31 (0.46)	0.25 (0.43)	3,4,6
Outcome variables from *UCDP* and *ACLED*					
State violence (ACLED)	7,330	5,513	0.14 (0.35)	0.08 (0.27)	–
State violence (UCDP)	7,330	5,513	0.05 (0.21)	0.03 (0.16)	–
Non-state violence (ACLED)	7,330	5,513	0.23 (0.42)	0.09 (0.29)	–
Non-state violence (UCDP)	7,330	5,513	0.05 (0.23)	0.01 (0.10)	–
One-sided violence (ACLED)	7,330	5,513	0.19 (0.39)	0.06 (0.24)	–
One-sided violence (UCDP)	7,330	5,513	0.06 (0.23)	0.02 (0.15)	–
Individual controls					
Urban	68,807	30,036	0.35 (0.48)	0.36 (0.48)	3,4,5,6
Age	68,807	30,036	36.24 (14.59)	38.28 (14.51)	3,4,5,6
Employment	68,807	30,036	0.40 (0.49)	0.27 (0.44)	3,4,5,6
Male	68,807	30,036	0.50 (0.50)	0.50 (0.50)	3,4,5,6
Country controls					
West Africa	68,807	30,036	0.26 (0.44)	0.85 (0.36)	–
East Africa	68,807	30,036	0.33 (0.47)	0.15 (0.36)	–
Former German colony	68,807	30,036	0.16 (0.36)	0.07 (0.25)	–
Landlocked	68,807	30,036	0.33 (0.47)	0.43 (0.49)	–
Ethnicity and district controls					
No. of slaves exported (per km^2)	68,807	30,036	1.31 (14.59)	5.44 (11.91)	–
Cities in 1800	68,807	30,036	0.11 (0.31)	0.16 (0.37)	–
Railway	68,807	30,036	0.47 (0.50)	0.23 (0.42)	–
Explorer	68,807	30,036	0.56 (0.50)	0.40 (0.49)	–
Missionary stations (per km^2)	68,807	30,036	0.30 (0.47)	0.12 (0.23)	–
Density of nighttime light	68,807	30,036	0.50 (0.62)	0.37 (0.44)	–
Distance to coast (1,000 km)	68,807	30,036	0.52 (0.32)	0.39 (0.36)	–
Distance to capital (1,000 km)	68,807	30,036	0.29 (0.22)	0.27 (0.21)	–
Fractionalization	68,807	30,036	0.12 (0.21)	0.12 (0.23)	3,4,5,6
Own ethnic share	68,807	30,036	0.88 (0.25)	0.88 (0.25)	3,4,5,6
No. of countries	12	9			

Notes: This table reports means and standard deviations of the variables by colonial status (anglophone vs. francophone). Standard deviations are in parentheses. The control variables are either at individual, ethnicity, district or country levels. Outcome variables on political violence (namely, state, non-state and one-sided violence) are at the 0.25*0.25 degree grid--cell level. The remaining outcome variables are at individual level. The last column lists the survey rounds for variables from the Afrobarometer data.

Table 2: National Identity and Colonial Status

	(1)	(2)	(3)	(4)
Anglophone	-0.12**	-0.17***	-0.18***	-0.14***
	(0.05)	(0.05)	(0.05)	(0.04)
Observations	98,843	98,843	98,843	98,843
Within-country R^2	0.003	0.003	0.005	0.005
Overall R^2	0.022	0.050	0.055	0.063
Geographic controls	–	Yes	Yes	Yes
Former German colony	–	Yes	Yes	Yes
Individual controls	–	–	Yes	Yes
Ethnic controls	–	–	–	Yes
District controls	–	–	–	Yes

Note: The dependent variable (*National identity*) measures the respondents' strength of national identification. *Anglophone* is a dummy for whether the respondent is from an anglophone country. All regressions include survey-round fixed effects. The geographic controls include indicators for region (western, southern, and eastern Africa) and 'landlockedness'. Former German colony is a dummy for whether the country was colonized by Germany prior to the First World War. Individual-level controls account for age, age squared, education level, religion, asset ownership, gender, employment status, and location (urban versus rural). Ethnicity-level controls account for urbanization levels in 1800, precolonial judicial hierarchy, access to colonial rail network, precolonial contact with European explorers, missionary activity during colonial times, exposure to slave trade, density of nighttime light, distance to the capital city, and distance to the coast. District-level controls are the share of the population in own ethnic group and ethnic fractionalization. Robust standard errors, two-way clustered at country and ethnicity level, are given in parentheses. **Significant at 5\%, ***significant at 1\%.

Table 3: National Identity and Colonial Status: RD Results for Western Africa

	(1)	(2)	(3)	(4)	(5)	(6)
Anglophone	-0.15***	-0.20***	-0.20***	-0.22***	-0.15**	-0.14***
	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)
Observations	43,013	12,748	12,748	12,748	12,748	12,748
R^2	0.045	0.075	0.076	0.084	0.097	0.102
Geographic controls	–	–	Yes	Yes	Yes	Yes
Former German colony	–	–	–	Yes	Yes	Yes
Individual controls	–	–	–	–	Yes	Yes
Ethnic controls	–	–	–	–	–	Yes
District controls	–	–	–	–	–	Yes

Note: Column (1) includes all respondents in western Africa. In columns (2)--(6), the observations are drawn from respondents who reside within 100 km of anglophone—francophone national borders in western Africa (i.e., the RD sample). *Distance to border* is the distance to either side of the nearest anglophone--francophone border. All regressions include survey-round fixed effects. *Legal origin controls* includes three dummies: respondents' trust in the judiciary, respondents' attitudes towards political freedom and respondents' attitudes towards press freedom. See Table 2 for descriptions of the remaining controls. Robust standard errors, two-way clustered at the country and ethnicity levels, are given in parentheses. **Significant at 5\%, *** significant at 1\%.

Table 4: National Identity and Colonial Status: RD Results for Cameroon

	(1)	(2)	(3)	(4)	(5)	(6)
Anglophone	-0.40***	-0.46***	-0.45***	-0.43***	-0.36***	-0.32***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.06)	(0.06)
Observations	2,230	880	880	880	880	880
R^2	0.087	0.213	0.218	0.251	0.296	0.324
Geographic controls	–	–	Yes	Yes	Yes	Yes
Former German colony	–	–	–	Yes	Yes	Yes
Individual controls	–	–	–	–	Yes	Yes
Ethnic controls	–	–	–	–	–	Yes
District controls	–	–	–	–	–	Yes

Note: Column (1) includes all respondents in Cameroon. In columns (2)--(6), the observations are drawn from the administrative regions near the anglophone--francophone border in Cameroon (see Figure 4). *Distance to border* is the distance to either side of the nearest anglophone--francophone border within Cameroon. All regressions include survey-round fixed effects. *Legal origin controls* includes three dummies: respondents' trust in the judiciary, respondents' attitudes towards political freedom and respondents' attitudes towards press freedom. See Table 2 for descriptions of the remaining controls. Robust standard errors, clustered at the ethnicity level, are given in parentheses. * Significant at 10%, ** significant at 5%, *** significant at 1%.

Table 5: Taxation, Security, and the Power of Chiefs---Results from All Observations

	Compliance Norm	Evasion Difficulty	Extortion Prevalence	Crime Prevalence	Chief Contact
	(1)	(2)	(3)	(4)	(5)
Anglophone	-0.02	-0.04***	0.12***	0.16***	0.10***
	(0.06)	(0.02)	(0.04)	(0.06)	(0.04)
Observations	58,703	54,101	28,045	80,954	69,850
Within-country R^2	0.014	0.003	0.005	0.010	0.074
Overall R^2	0.028	0.019	0.030	0.032	0.111

Note: The dependent variables are listed at the top of each column. All regressions include survey-round fixed effects, geographic controls, the indicator for a Germany colony, and controls at individual, ethnicity, and district levels (see Table 2 for descriptions of these controls). Standard errors, two-way clustered at the ethnicity and country levels, are given in parentheses.*** Significant at 1%.

Table 6: Taxation, Security, and the Power of Chiefs: RD Results for Western Africa

	Compliance Norm	Evasion Difficulty	Extortion Prevalence	Crime Prevalence	Chief Contact
	(1)	(2)	(3)	(4)	(5)
			Panel A: All observations in West Africa		
Anglophone	-0.02	-0.08***	0.09***	0.06	0.09***
	(0.07)	(0.02)	(0.03)	(0.06)	(0.04)
Observations	28,214	25,937	13,316	36,398	29,440
R^2	0.024	0.017	0.075	0.041	0.104
			Panel B: RD sample		
Anglophone	-0.04	-0.10	0.08	0.00	0.10
	(0.08)	(0.03)	(0.03)	(0.06)	(0.03)
Observations	10,234	9,372	4,730	11,726	7,900
R^2	0.044	0.024	0.076	0.040	0.139

Note: The outcome variables are listed at the top of each column. In Panel B, the observations are drawn from respondents residing within 100 km of borders between anglophone and francophone countries in western Africa. All regressions include survey-round fixed effects and the remaining controls at individual, ethnicity, and district levels (see Table 2). The RD regressions (Panel B) include the RD distance to the border (see Table 3) and border fixed effects. Robust standard errors, two-way clustered at the ethnicity and country levels, are given in parentheses.*** Significant at 1%.

Table 7: Taxation, Security, and the Power of Chiefs: RD Results for Cameroon

	Compliance	Evasion	Extortion	Crime	Chief
	Norm	Difficulty	Prevalence	Prevalence	Contact
	(1)	(2)	(3)	(4)	(5)
Anglophone	0.02	0.02	0.17***	0.31***	-0.00
	(0.06)	(0.05)	(0.08)	(0.13)	(0.10)
Observations	855	767	446	914	445
R^2	0.098	0.052	0.113	0.066	0.201

Note: The outcome variables are listed at the top of each column. The observations are drawn from administrative regions near the anglophone--francophone border in Cameroon (see Figure 4). All regressions include survey-round fixed effects, the RD distance control (see Table 3), and the remaining controls at individual, ethnicity, and district levels (see Table 2). Robust standard errors, clustered at the ethnicity level, are given in parentheses.** Significant at 5%.

Table 8: Colonial Status and Conflict

	State violence		Nonstate violence		One-sided violence	
	UCDP	ACLED	UCDP	ACLED	UCDP	ACLED
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: All Observations					
Anglophone	0.11	0.09	0.11	0.21	0.10	0.21
	(0.05)	(0.04)	(0.02)	(0.05)	(0.05)	(0.06)
N = 12,843						
Within-country R^2	0.024	0.055	0.014	0.050	0.021	0.051
Overall R^2	0.081	0.139	0.123	0.209	0.084	0.197
	Panel A: Western Africa					
Anglophone	0.12	0.11	0.13	0.24	0.10	0.24
	(0.05)	(0.05)	(0.02)	(0.05)	(0.05)	(0.06)
N = 6,193						
Overall R^2	0.082	0.140	0.148	0.255	0.107	0.241
	Panel A: RD Sample					
Anglophone	0.14	0.04	0.06	0.23	0.09	0.23
	(0.06)	(0.04)	(0.01)	(0.06)	(0.07)	(0.05)
N = 808						
Overall R^2	0.326	0.085	0.101	0.227	0.379	0.290

Note: The units of observations are 0.25*0.25 degree cells. *Anglophone* is an indicator variable for whether the cell lies in an anglophone territory. The dependent variables, listed at the top of each column, are constructed from two different data sets (UCDP and ACLED); these variables are indicators for whether or not a particular type of conflict (state violence, nonstate violence, or one-sided violence) was observed in each cell over the sample period (1989--2016 for UCDP, 1997--2016 for ACLED). Results are reported for three samples. Panel A includes all cells, and Panel B includes just the cells in western Africa. All of the location-level controls (e.g. ethnicity-level controls; see Table 2) are included by assigning to each grid cell the values of the geographic unit to which that cell belongs. Distance controls (distance to the capital city and to the coast) are reconstructed at the grid level. Panel C reports RD results using cells within 100 km of the anglophone--francophone borders in western Africa. The RD results include additional controls for distance to the nearest border as well as border fixed effects. Robust standard errors, clustered at the ethnicity and country levels, are given in parentheses. * Significant at 10%, ** significant at 5%, *** significant at 1%.

Figure 1: Anglophone and Francophone Countries in the Data Set

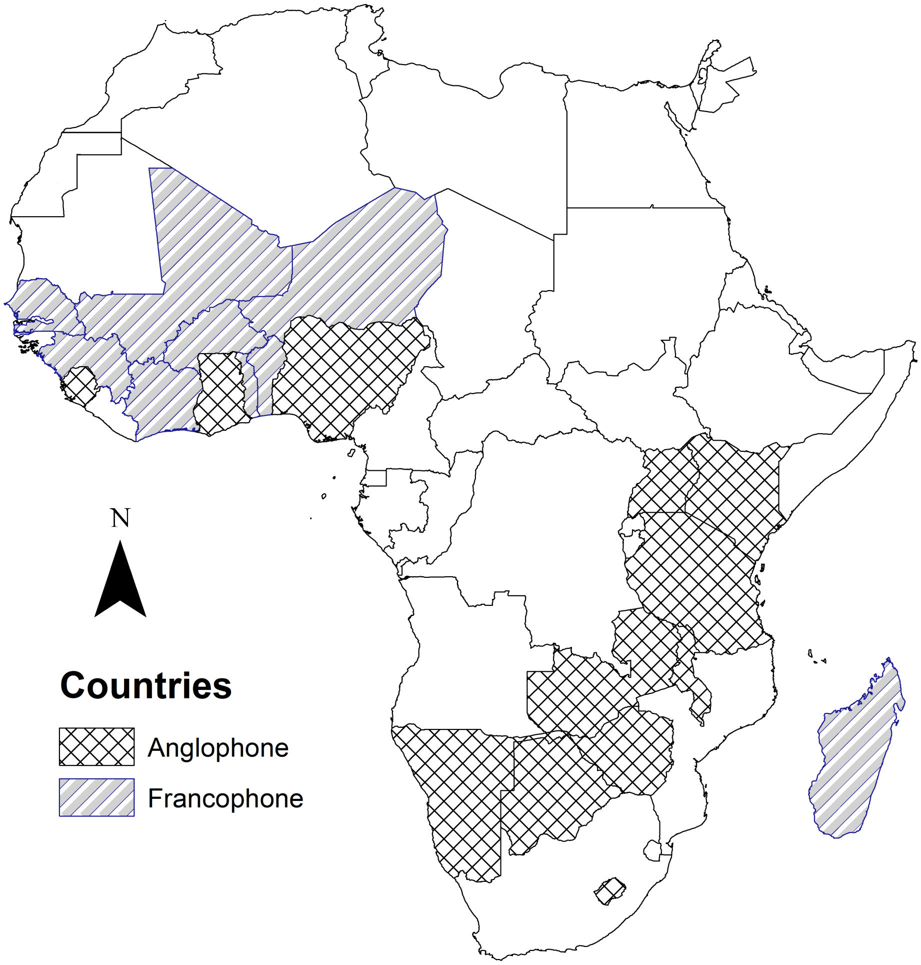
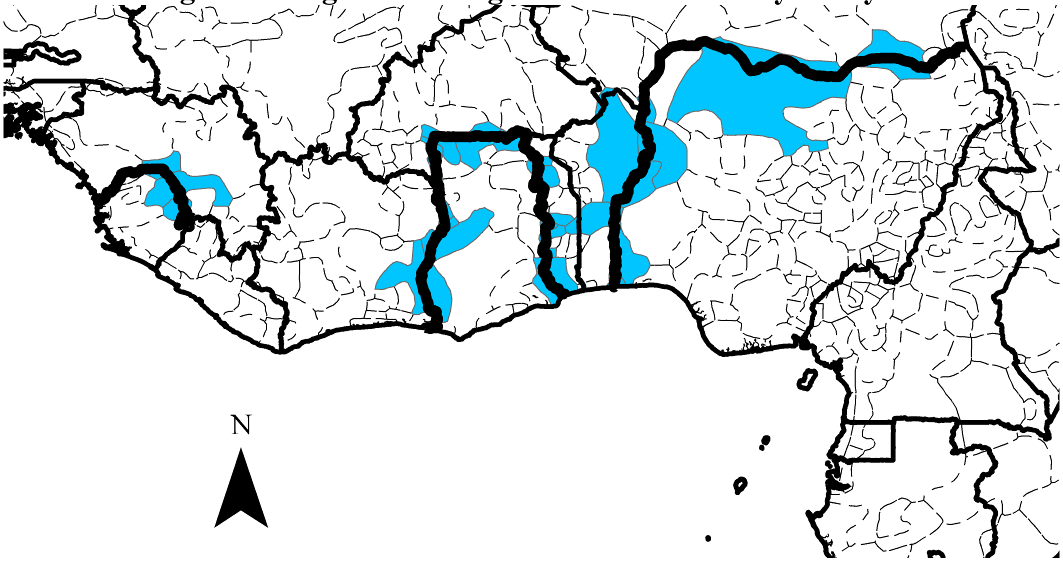
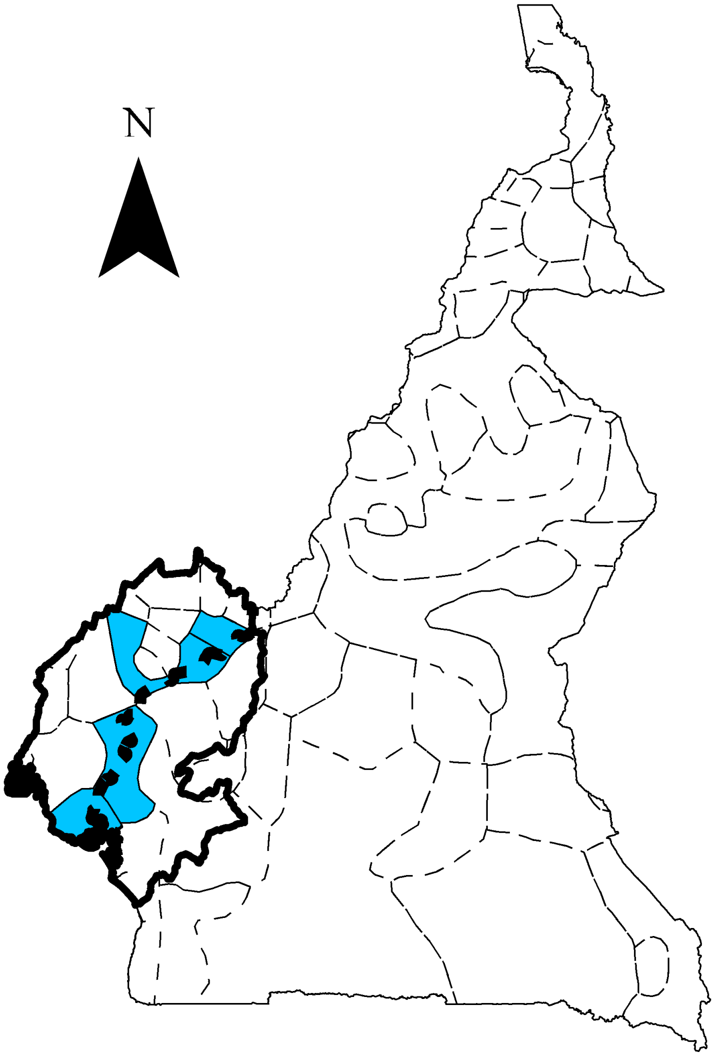


Figure 2: Regions for Regression Discontinuity Analysis



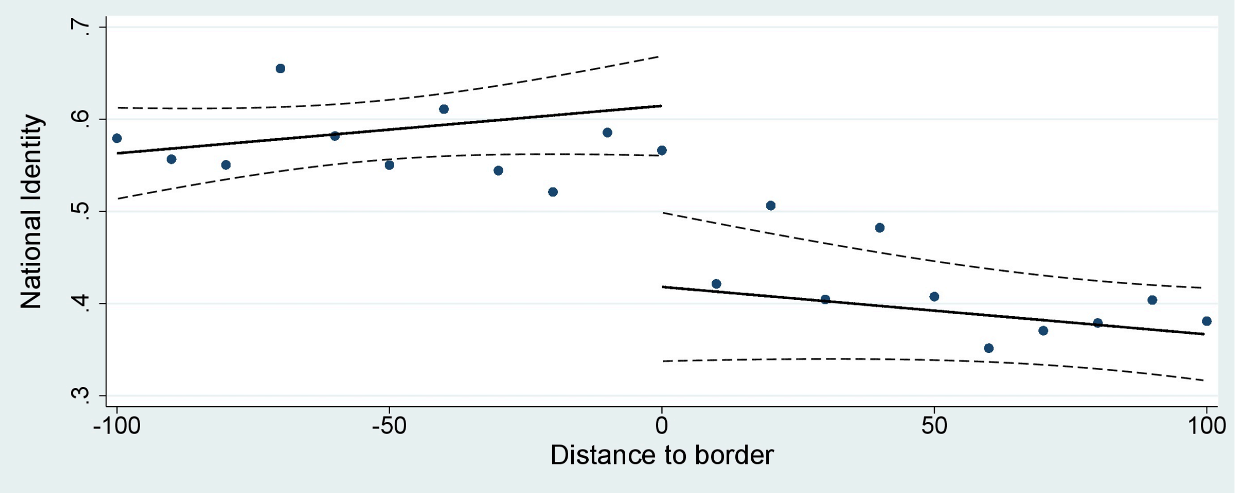
Note: The solid lines in this map represent national borders in western Africa, with the anglophone-francophone borders represented by the thickest lines. The thin broken lines mark the borders of ethnic groups' historical homelands. The shaded areas are historical homelands of ethnic groups that were split between anglophone and francophone countries.

Figure 3: Map of Cameroon



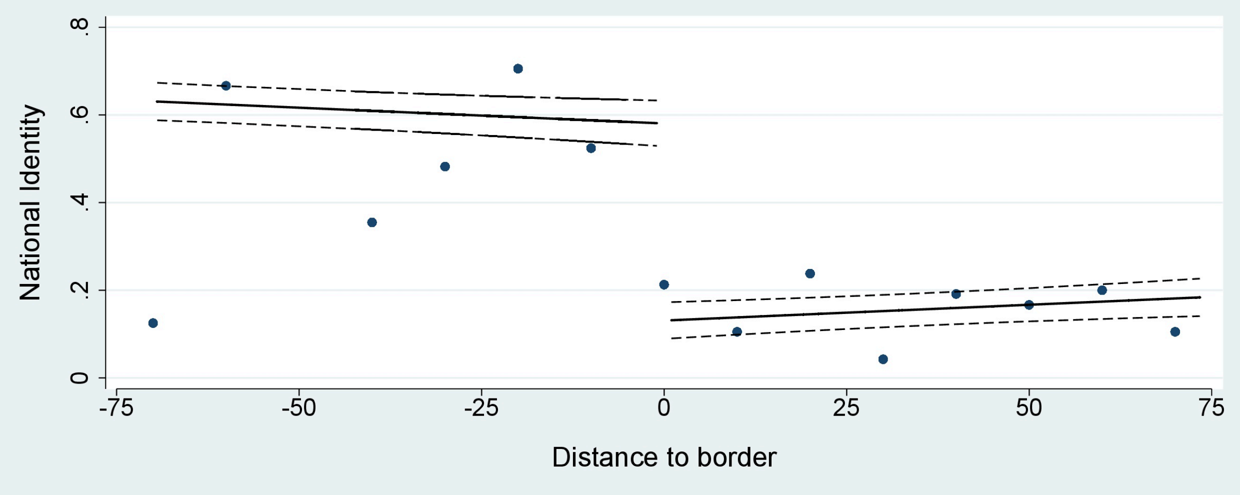
Note: The area circumscribed by the thick solid line represents the region from which observations for our RD analyses were drawn. The heavy broken line (within that region) marks the anglophone{francophone border, where the region west (resp. east) of that line is anglophone (resp. francophone). As before, the thinnest lines indicate borders of ethnic groups' historical homelands; the shaded areas are historical homelands of ethnic groups that were split between anglophone and francophone Cameroon.

Figure 4: National Identity in Western Africa, by Distance to Anglophone—Francophone Border



Note: The Figure shows—by distance to border in km—local averages in 10 km bins. The distance from the francophone—anglophone border increases as we move away from the center point (0). Negative/positive values represent distance into francophone/anglophone territories (from the border).

Figure 5: National Identity in Cameroon, by Distance to Anglophone—Francophone Border



Note: The Figure shows—by distance to border in km—local averages in 10 km bins. The distance from the francophone—anglophone border increases as we move away from the center point (0). Negative/positive values represent distance into francophone/anglophone territories (from the border).

Appendix

A. Variables and Data Sources

Afrobarometer Variables

All of the Afrobarometer data are downloaded from the official Afrobarometer website: http://www.afrobarometer.org/data/merged-data

National identity. An individual-level binary index indicating respondents' attitudes towards national versus ethnic identity; see Section 2. Survey questions: Q82 from round 3, Q83 from round 4, Q85B from round 5, and Q88B from round 6.

Compliance norm. An individual-level binary index reflecting respondents' views regarding tax evasion; see Section 3. Survey questions: Q76B from round 5 and Q75B from round 6.

Evasion difficulty. An individual-level binary index that measures respondents' views about the difficulty of evading taxes; see Section 3. Survey questions: Q75C from rounds 5 and 6.

Extortion prevalence. An individual-level binary index indicating the prevalence of extortion activity by non-state actors; see Section 3. Survey question: Q74 from round 5.

Crime prevalence. An individual-level count variable that measures respondents' experience with crime incidents; see Section 3. Survey questions: Q9A-C from round 4, Q9B and Q10A-B from round 5, and Q10B and Q11A-B from round 6.

Chief contact. An individual-level binary index measuring how frequently the respondents contacted their local chiefs; see Section 3. Survey questions: Q32F from round 3, Q27B from round 4, and Q24E from round 6.

Urban. An individual-level indicator set equal to 1 if the respondent is from an urban area (and set to 0 otherwise). Survey questions: URBRUR from rounds 3–6.

Age. Age of the respondent, ranging from 18 to 105. Survey questions: Q1 from rounds 3–6.

Employment. Employment status of the respondent, set equal to 1 if the respondent is employed (either full-time or part-time) and otherwise set equal to 0. Survey questions: Q94 from rounds 3 and 4, Q96 from round 5, and Q97 from round 6.

Education. Dummies for the respondents' level of education attainment based on nine education attainment groups. Survey questions: Q90 from round 3, Q89 from round 4, and Q97 from rounds 5 and 6.

Religion. Dummies for eight religion groups. Survey questions: Q98A from rounds 3–6.

Gender. An indicator variable for the respondent's gender. Survey questions: Q101 from rounds 3–6.
Wealth. Three dummies for the ownership of a radio, a television, or an automobile. Survey questions: Q92A–C from rounds 3 and 4 and Q90A–C from rounds 5 and 6.

Trust in courts of law. Measures the respondents' level of trust in courts of law; see Section 2.3. Survey questions: Q49H from round 3, Q55I from round 4, and Q59J from round 5.

Political freedom to organise Indicator of respondents' attitudes towards political freedom to organise; see Section 2.3. Survey questions: Q25 from round 3 and Q19 from rounds 4 and 5.

Press freedom. Indicator of respondents' attitudes towards press freedom; see Section 2.3. Survey questions: Q26 from round 3 and Q20 from rounds 4 and 5.

Other Variables

Anglophone. An indicator for whether (or not) the observation is from an anglophone country.

Region indicators. Eastern Africa includes Tanzania, Kenya, Uganda, and Madagascar; western Africa includes Benin, Burkina Faso, Cote d'Ivoire, Ghana, Guinea, Mali, Niger, Nigeria, Senegal, Sierra Leone, and Togo; southern Africa includes Malawi, Zambia, and Zimbabwe. This categorisation follows Bratton and van de Walle (1997).

Landlocked. A binary indicator set equal to 1 if the country is landlocked (and set to 0 otherwise).

Former German colony. A dummy variable for Tanzania, Namibia, and Togo— which were German colonies prior to the First World War.

Share of own ethnic group. A district-level index ranging from 0 to 1; it measures the share of the district's population that is of same ethnicity as the respondent. This index is calculated (from Afrobarometer data) following Nunn and Wantchekon (2011).

Ethnic fractionalisation. A district-level index ranging from 0 to 1; it measures the probability that two randomly selected individuals from a district belong to different ethnic groups. This index is calculated (from Afrobarometer data) following Alesina et al. (2003).

Slave export. Total slave export count, from both trans-Atlantic and Indian trade, for each ethnic group. Source: Nunn and Wantchekon (2011).

Cities in 1800. An indicator for whether (or not) the focal ethnic group's historical homeland contained a city populated by at least 20,000 inhabitants in 1800. Source: Chandler and Fox (1974).

Historical homelands of ethnic groups. Provided by the digital version of Mur- dock's (1959) Ethnolinguistic Map. Land area of each ethnic homeland is computed using the 'shapefile' from Nunn and Wantchekon (2011).

Railway indicator. A dummy variable for whether (or not) there was a colonial railway station within the focal ethnic group's historical homeland. Source: Nunn and Wantchekon (2011).

European explorers. An indicator variable for whether (or not) European explorers passed through the focal ethnic group's historical homeland during the precolonial era. Source: Nunn and Wantchekon (2011).

Missionary activity. The number of mission stations located in the focal ethnic group's historical homeland. Source: Nunn (2010).

Light density. Average of night-time light density per square kilometre within the focal ethnic group's historical homeland. Source: https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Judicial hierarchy. The number of jurisdictional hierarchies beyond the local community. Sources: Murdock (1967) and Nunn and Wantchekon (2011).

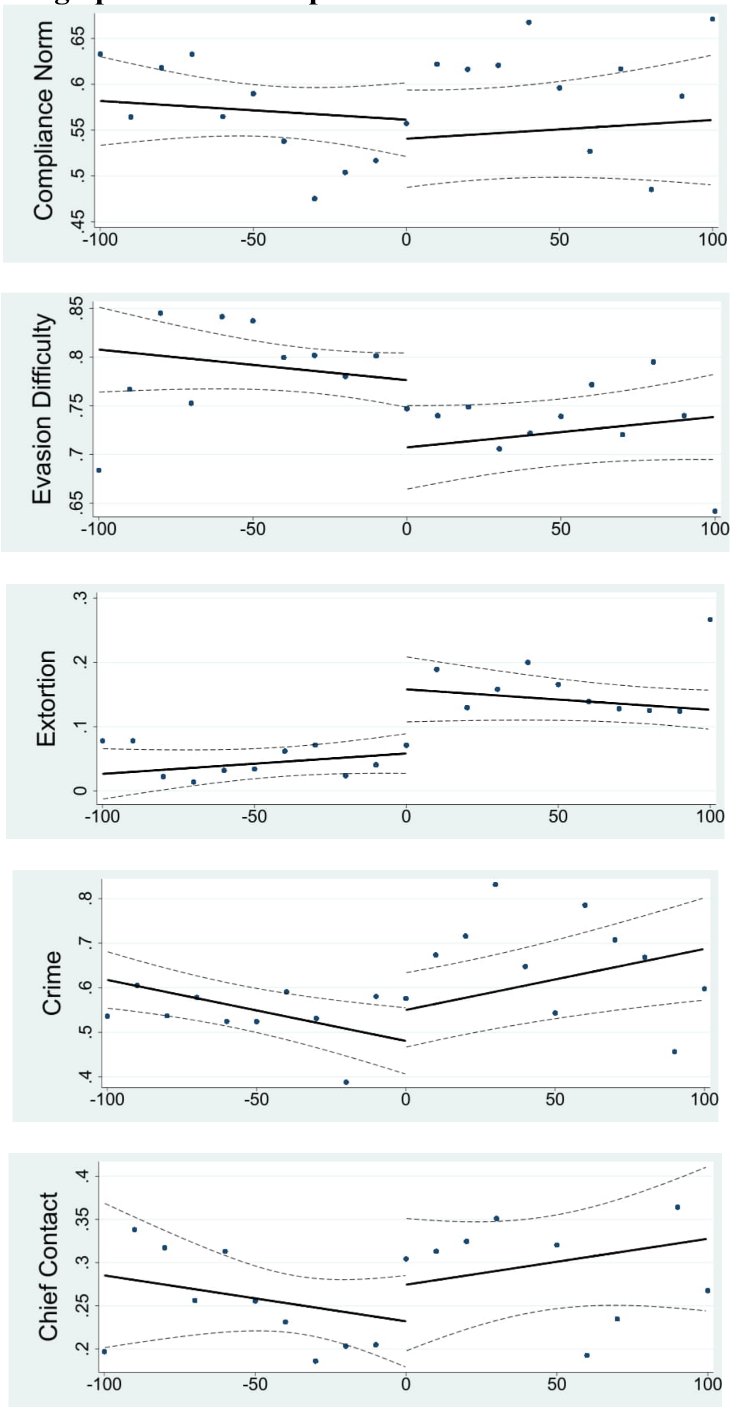
Distance to capital city. Distance between (the centroid of) each ethnic homeland and the capital city. Data on capital cities are from the Natural Earth database: http://www.naturalearthdata.com/downloads/10m-cultural-vectors/

Distance to coast. The distance between the centroid of the focal ethnic homeland and the nearest coast.

Conflict indicators. Three dummy variables—State violence, Non-state violence, and One-sided violence—are constructed, at the (0.25 0.25)-degree–cell level, using data from ACLED and UCDP. Each of these conflict variables indicates whether the respective type of violence occurred in each cell over the period covered by the two data sets; see Section 3.2. Sources: for UCDP, Sundberg and Melander (2013) and Croicu and Sundberg (2017); for ACLED, Raleigh et al. (2010) and Raleigh and Dowd (2017).

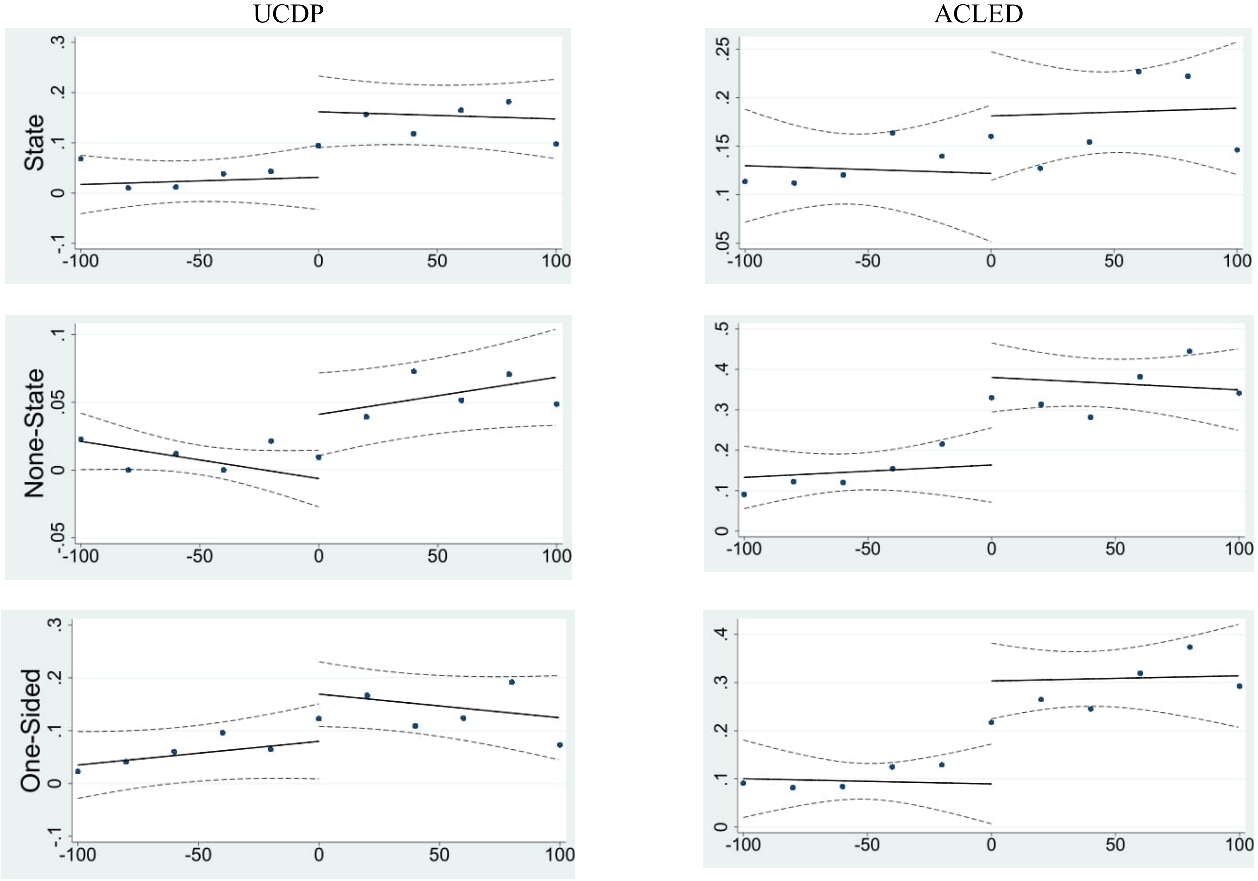
B. Figures

Figure B.1: RD Plots for Taxation, Security and the Power of Chiefs, by Distance to Anglophone—Francophone Borders in West Africa



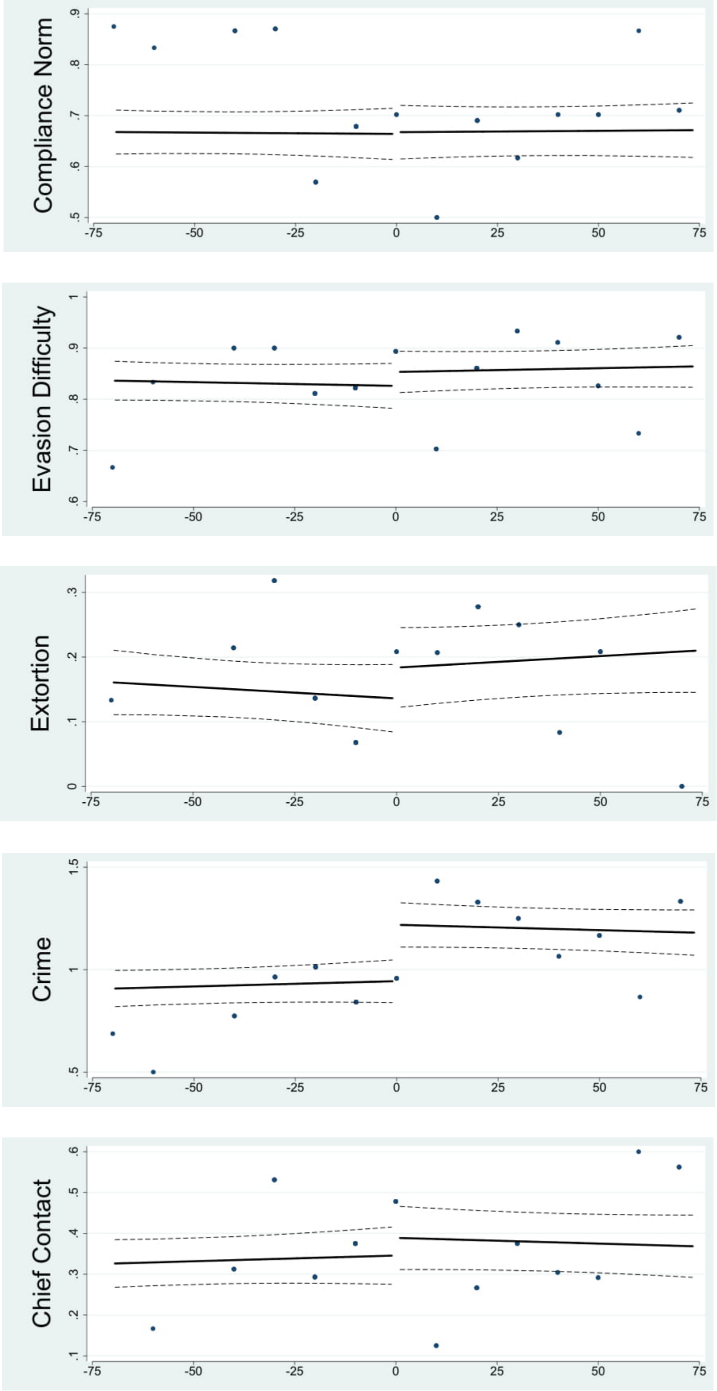
Note: The Figure shows—by distance to border in km—local averages in 10 km bins. The distance from the francophone—anglophone border increases as we move away from the center point (0). Negative/positive values represent distance into francophone/anglophone territories (from the border).

Figure B.2: State Violence, Non-State Violence and One-Sided Violence from UCDP and ACLED Data Sets, by Distance to the Anglophone—Francophone Border in West Africa.



Note: The Figure shows—by distance to border in km—local averages in 20 km bins. The distance from the francophone—anglophone border increases as we move away from the center point (0). Negative/positive values represent distance into francophone/anglophone territories (from the border).

Figure B.3: RD Plots for Taxation, Security and the Power of Chiefs, by Distance to Anglophone—Francophone Borders in Cameroon



Note: The Figure shows—by distance to border in km—local averages in 10 km bins. The distance from the francophone—anglophone border increases as we move away from the center point (0). Negative/positive values represent distance into francophone/anglophone territories (from the border).

VITA

NAME OF AUTHOR: Boqian Jiang

PLACE OF BIRTH: Guiyang, China

DATE OF BIRTH: May, 1989

EDUCATION:

2011	B.A., Southwestern University of Finance and Economics, Chengdu, China
2013	M.A., Toulouse School of Economics, Toulouse, France

RESEARCH EXPERIENCE AND EMPLOYMENT:

2014 – 2018	Research Assistant, Department of Economics, Syracuse University
2015 – 2018	Research Associate, Center for Policy Research, Syracuse University

AWARDS AND HONORS:

2010	Bank of China Scholarship, Southwestern University of Finance and Economics
2013 – 2018	Syracuse University Graduate Assistantship, Syracuse University