MIT 14.582 PhD International Economics II — Lectures 12-13: Economic Geography and Urban Economics (Empirics) —

Dave Donaldson

Spring 2025

Empirical work on the causes of agglomeration

- Last lecture talked about reasons we might expect to see a location's Y/L increasing in its L
  - But is this true? And how strong is this agglomeration effect?
  - That will be the goal of today's and next lecture
  - Admittedly, like most of the literature, we won't have much to say about mechanisms for *why* agglomeration happens

- Some surveys on this theme:
  - Rosenthal and Strange (2004, *Handbook of Urban and Regional Econ*)
  - Combes and Gobillon (2015, *Handbook of Urban and Regional Econ*)

- Broadly, three approaches:
  1. Estimating agglomeration economies "directly"
  2. Estimating agglomeration economies from the extent of agglomeration in an observed spatial equilibrium. (Can think of EG (1997) like this.)
  3. Testing for multiple equilibria and/or multiple steady-states (which are often a consequence of positive agglomeration economies)

- For now, our main focus will be #1

Estimating agglomeration economies directly

- A large literature has argued that if agglomeration economies exist then units of production (and factors) should be more productive if they are surrounded by other producers
- Three nice examples:
  - Henderson (2003, JUE) on across-firm (within-location) externalities
  - Moretti (2004, AER) on local (within-city) human capital externalities
  - Arzaghi and Henderson (2008, REStud) on Manhattan's advertising industry
- A central challenge with this approach is an analogy to the challenge that faces the 'peer effects' literature (e.g. Manski, 1993): does one unit actually affect a proximate unit, or are proximate units just similar on unobservable dimensions?
- We will discuss three recent studies that offer natural experiment approaches to this question.
  - Greenstone, Hornbeck and Moretti (JPE, 2010)
  - Kline and Moretti (QJE, 2014)
  - Peters (ECMA, 2022)

Greenstone, Hornbeck and Moretti (2010)

- GHM look at the effect that 'million dollar plants' (huge industrial plants) have on incumbent firms in the vicinity of the new MDP
- Consider the following example (from paper):
  - BMW did worldwide search for new plant location in 1991. 250 locations narrowed to 20 US counties. Then announced 2 finalists: Omaha, NB and Greenville-Spartanburg, SC. Finally, chose Greenville-Spartanburg.
  - Why? BMW says:
    - Low costs of production: low union density, supply of quality workers, numerous global firms in area (including 58 German companies), good transport infrastructure (rail, air, highway, port access), and access to key local services.
    - Subsidy ($115 million) received from local government.
- GHM obtain list of the winner and loser counties for 82 MDP openings and compare winners to losers (rather than comparing winners to all 3,000 other counties, or to counties that look similar on observables).

Greenstone, Hornbeck and Moretti (2010)

TABLE 3

County and Plant Characteristics by Winner Status, 1 Year Prior to a Million Dollar Plant Opening

	All Plants					Within Same Industry (Two-Digit SIC)				
	Winning Counties (1)	Losing Counties (2)	All U.S. Counties (3)	*t*-Statistic (Col. 1 − Col. 2) (4)	*t*-Statistic (Col. 1 − Col. 3) (5)	Winning Counties (6)	Losing Counties (7)	All U.S. Counties (8)	*t*-Statistic (Col. 6 − Col. 7) (9)	*t*-Statistic (Col. 6 − Col. 8) (10)
A. County Characteristics										
No. of counties	47	73				16	19			
Total per capita earnings ($)	17,418	20,628	11,259	**−2.05**	**5.79**	20,230	20,528	11,378	−.11	**4.62**
% change, over last 6 years	.074	.096	.037	−.81	1.67	.076	.089	.057	−.28	.57
Population	322,745	447,876	82,381	−1.61	**4.33**	357,955	504,342	83,430	−1.17	**3.26**
% change, over last 6 years	.102	.051	.036	**2.06**	**3.22**	.070	.032	.031	1.18	1.63
Employment-population ratio	.535	.579	.461	−1.41	**3.49**	.602	.569	.467	.64	**3.63**
Change, over last 6 years	.041	.047	.023	−.68	**2.54**	.045	.038	.028	.39	1.57
Manufacturing labor share	.314	.251	.252	**2.35**	**3.12**	.296	.227	.251	1.60	1.17
Change, over last 6 years	−.014	−.031	−.008	1.52	−.64	−.030	−.040	−.007	.87	**−3.17**
B. Plant Characteristics										
No. of sample plants	18.8	25.6	7.98	−1.35	**3.02**	2.75	3.92	2.38	−1.14	.70
Output ($1,000s)	190,039	181,454	123,187	.25	**2.14**	217,950	178,958	132,571	.41	1.25
% change, over last 6 years	.082	.082	.118	.01	−.97	−.061	.177	.182	−1.23	**−3.38**
Hours of labor (1,000s)	1,508	1,168	877	1.52	**2.43**	1,738	1,198	1,050	.92	1.33
% change, over last 6 years	.122	.081	.115	.81	.14	.160	.023	.144	.85	.13

Note.—For each case to be weighted equally, counties are weighted by the inverse of their number per case. Similarly, plants are weighted by the inverse of their number per county multiplied by the inverse of the number of counties per case. The sample includes all plants reporting data in the ASM for each year between the MDP opening and 8 years prior. Excluded are all plants owned by the firm opening an MDP. Also excluded are all plants from two uncommon two-digit SIC values so that subsequently estimated clustered variance matrices would always be positive definite. The sample of all U.S. counties excludes winning counties and counties with no manufacturing plant reporting data in the ASM for 9 consecutive years. These other U.S. counties are given equal weight within years and are weighted across years to represent the years of MDP openings. Reported *t*-statistics are calculated from standard errors clustered at the county level. *t*-statistics greater than 2 are reported in bold. All monetary amounts are in 2006 U.S. dollars.

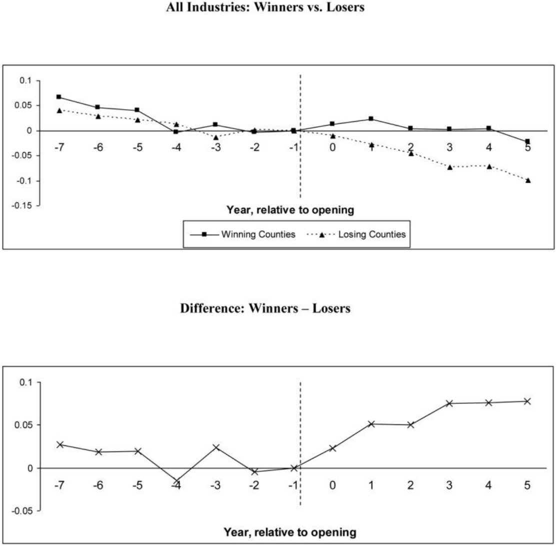


FIG. 1.—All incumbent plants' productivity in winning versus losing counties, relative to the year of an MDP opening. These figures accompany table 4.

Greenstone, Hornbeck and Moretti (2010)

TABLE 5

Changes in Incumbent Plant Productivity Following an MDP Opening

	All Counties: MDP Winners − MDP Losers (1)	All Counties: MDP Winners − MDP Losers (2)	MDP Counties: MDP Winners − MDP Losers (3)	MDP Counties: MDP Winners − MDP Losers (4)	All Counties: Random Winners (5)
			A. Model 1		
Mean shift	.0442*	.0435*	.0524**	.0477**	− 0.0496***
	(.0233)	(.0235)	(.0225)	(.0231)	(.0174)
				[$170 m]	
R^2	.9811	.9812	.9812	.9860	~0.98
Observations (plant by year)	418,064	418,064	50,842	28,732	~400,000
			B. Model 2		
Effect after 5 years	.1301**	.1324**	.1355***	.1203**	−.0296
	(.0533)	(.0529)	(.0477)	(.0517)	(.0434)
				[$429 m]	
Level change	.0277	.0251	.0255	.0290	.0073
	(.0241)	(.0221)	(.0186)	(.0210)	(.0223)
Trend break	.0171*	.0179**	.0183**	.0152*	− 0.0062
	(.0091)	(.0088)	(.0078)	(.0079)	(.0063)
Pre-trend	−.0057	−.0058	−.0048	−.0044	−.0048
	(.0046)	(.0046)	(.0046)	(.0044)	(.0040)
R^2	.9811	.9812	.9813	.9861	~.98
Observations (plant by year)	418,064	418,064	50,842	28,732	~400,000
Plant and industry by year fixed effects	Yes	Yes	Yes	Yes	Yes
Case fixed effects	No	Yes	Yes	Yes	NA
Years included	All	All	All	$-7 \leq \tau \leq 5$	All

Note.—The table reports results from fitting several versions of eq. (8). Specifically, entries are from a regression of the natural log of output on the natural log of inputs, year by two-digit SIC fixed effects, plant fixed effects, and case fixed effects. In model 1, two additional dummy variables are included for whether the plant is in a winning county 7 to 1 years before the MDP opening or 0 to 5 years after. The reported mean shift indicates the difference in these two coefficients, i.e., the average change in TFP following the opening. In model 2, the same two dummy variables are included along with pre- and post-trend variables. The shift in level and trend are reported, along with the pre-trend and the total effect evaluated after 5 years. In cols. 1, 2, and 5, the sample is composed of all manufacturing plants in the ASM that report data for 14 consecutive years, excluding all plants owned by the MDP firm. In these models, additional control variables are included for the event years outside the range from $\tau = -7$ through $\tau = 5$ (i.e., −20 to −8 and 6 to 17). Column 2 adds the case fixed effects that equal one during the period that τ ranges from −7 through 5. In cols. 3 and 4, the sample is restricted to include only plants in counties that won or lost an MDP. This forces the industry by year fixed effects to be estimated solely from plants in these counties. For col. 4, the sample is restricted further to include only plant by year observations within the period of interest (where τ ranges from −7 to 5). This forces the industry by year fixed effects to be estimated solely on plant by year observations that identify the parameters of interest. In col. 5, a set of 47 plant openings in the entire country were randomly chosen from the ASM in the same years and industries as the MDP openings (this procedure was run 1,000 times, and reported are the means and standard deviations of those estimates). For all regressions, plant by year observations are weighted by the plant's total value of shipments 8 years prior to the opening. Plants not in a winning or losing county are weighted by their total value of shipments in that year. All plants from two uncommon two-digit SIC values were excluded so that estimated clustered variance-covariance matrices would always be positive definite. In brackets is the value in 2006 U.S. dollars from the estimated increase in productivity: the percentage increase is multiplied by the total value of output for the affected incumbent plants in the winning counties. Reported in parentheses are standard errors clustered at the county level.

* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

Greenstone, Hornbeck and Moretti (2010)

TABLE 6
CHANGES IN INCUMBENT PLANT OUTPUT AND INPUTS FOLLOWING AN MDP OPENING

	Output (1)	Worker Hours (2)	Machinery Capital (3)	Building Capital (4)	Materials (5)
Model 1: mean shift	.1200***	.0789**	.0401	.1327*	.0911***
	(.0354)	(.0357)	(.0348)	(.0691)	(.0302)
Model 2: after 5 years	.0826*	.0562	−.0089	−.0077	.0509
	(.0478)	(.0469)	(.0300)	(.0375)	(.0541)

NOTE.—The table reports results from fitting versions of eq. (8) for each of the indicated outcome variables (in logs). See the text for more details. Standard errors clustered at the county level are reported in parentheses.

* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

Greenstone, Hornbeck and Moretti (2010)

TABLE 7

CHANGES IN INCUMBENT PLANT PRODUCTIVITY FOLLOWING AN MDP OPENING FOR INCUMBENT PLANTS IN THE MDP'S TWO-DIGIT INDUSTRY AND ALL OTHER INDUSTRIES

	All Industries (1)	MDP's Two-Digit Industry (2)	All Other Two-Digit Industries (3)
		A. Model 1	
Mean shift	.0477**	.1700**	.0326
	(.0231)	(.0743)	(.0253)
	[$170 m]	[$102 m]	[$104 m]
R^2	.9860	.9861	
Observations	28,732	28,732	
		B. Model 2	
Effect after 5 years	.1203**	.3289	.0889*
	(.0517)	(.2684)	(.0504)
	[$429 m]	[$197 m]	[$283 m]
Level change	.0290	.2814***	.0004
	(.0210)	(.0895)	(.0171)
Trend break	.0152*	.0079	.0147*
	(.0079)	(.0344)	(.0081)
Pre-trend	−.0044	−.0174	−.0026
	(.0044)	(.0265)	(.0036)
R^2	.9861	.9862	
Observations	28,732	28,732	

NOTE.—The table reports results from fitting versions of eq. (8). As a basis for comparison, col. 1 reports estimates from the baseline specification for incumbent plants in all industries (baseline estimates for incumbent plants in all industries, col. 4 of table 5). Columns 2 and 3 report estimates from a single regression, which fully interacts the winner/loser and pre/post variables with indicators for whether the incumbent plant is in the same two-digit industry as the MDP or a different industry. Reported in parentheses are standard errors clustered at the county level. The numbers in brackets are the value (2006 U.S. dollars) from the estimated increase in productivity: the percentage increase is multiplied by the total value of output for the affected incumbent plants in the winning counties.

* Significant at the 10 percent level.

** Significant at the 5 percent level.

*** Significant at the 1 percent level.

Greenstone, Hornbeck and Moretti (2010)

TABLE 8

Changes in Incumbent Plant Productivity Following an MDP Opening, by Measures of Economic Distance between the MDP's Industry and Incumbent Plant's Industry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CPS worker transitions	.0701***						.0374
	(.0237)						(.0260)
Citation pattern		.0545***					.0256
		(.0192)					(.0208)
Technology input			.0320*				.0501
			(.0173)				(.0421)
Technology output				.0596***			.0004
				(.0216)			(.0434)
Manufacturing input					.0060		−.0473
					(.0123)		(.0289)
Manufacturing output						.0150	−.0145
						(.0196)	(.0230)
R^2	.9852	.9852	.9851	.9852	.9851	.9852	.9853
Observations	23,397	23,397	23,397	23,397	23,397	23,397	23,397

Note.—The table reports results from fitting versions of eq. (9), which is modified from eq. (8). Building on the model 1 specification in col. 4 of table 5, each column adds interaction terms between winner/loser and pre/post status with the indicated measures of how an incumbent plant's industry is linked to its associated MDP's industry (a continuous version of results in table 7). These industry linkage measures are defined and described in table 2, and here the measures are normalized to have a mean of zero and a standard deviation of one. The sample of plants is that in col. 4 of table 5, but it is restricted to plants that have industry linkage data for each measure. For assigning this linkage measure, the incumbent plant's industry is held fixed at its industry the year prior to the MDP opening. Whenever a plant is a winner or loser more than once, it receives an additive dummy variable and interaction term for each occurrence. Reported in parentheses are standard errors clustered at the county level.

* Significant at the 10 percent level.

** Significant at the 5 percent level.

*** Significant at the 1 percent level.

Greenstone, Hornbeck and Moretti (2010)

TABLE 9
CHANGES IN COUNTIES' NUMBER OF PLANTS, TOTAL OUTPUT, AND SKILL-ADJUSTED WAGES FOLLOWING AN MDP OPENING

	A. CENSUS OF MANUFACTURES		B. CENSUS OF POPULATION
	Dependent Variable: Log(Plants) (1)	Dependent Variable: Log(Total Output) (2)	Dependent Variable: Log(Wage) (3)
Difference-in-difference	.1255**	.1454	.0268*
	(.0550)	(.0900)	(.0139)
R^2	.9984	.9931	.3623
Observations	209	209	1,057,999

NOTE.—The table reports results from fitting three regressions. In panel A, the dependent variables are the log of number of establishments and the log of total manufacturing output in the county, based on data from the Census of Manufactures. Controls include county, year, and case fixed effects. Reported are the county-level difference-in-difference estimates for receiving an MDP opening. Because data are available every 5 years, depending on the census year relative to the MDP opening, the sample years are defined to be 1–5 years before the MDP opening and 4–8 years after the MDP opening. Thus, each MDP opening is associated with one earlier date and one later date. The col. 1 model is weighted by the number of plants in the county in years −6 to −10, and the col. 2 model is weighted by the county's total manufacturing output in years −6 to −10. In panel B, the dependent variable is log wage and controls include dummies for age by year, age squared by year, education by year, sex by race by Hispanic by citizen, and case fixed effects. Reported is the county-level difference-in-difference estimate for receiving an MDP opening. Because data are available every 10 years, the sample years are defined to be 1–10 years before the MDP opening and 3–12 years after the MDP opening. As in panel A, each MDP opening is associated with one earlier date and one later date. The sample is restricted to individuals who worked more than 26 weeks in the previous year, usually work more than 20 hours per week, are not in school, are at work, and work for wages in the private sector. The number of observations reported refers to unique individuals: some Integrated Public Use Microdata Series county groups include more than one Federal Information Processing Standard (FIPS), so all individuals in a county group were matched to each potential FIPS. The same individual may then appear in more than one FIPS, and observations are weighted to give each unique individual the same weight. Reported in parentheses are standard errors clustered at the county level.

* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

How to interpret the magnitude of GHM's estimates?

- A common functional form approach (consistent with the Dixit-Stiglitz microfoundation you have seen elsewhere in 14.582) for agglomeration effects in production is (where $\bar{A}_i$ is given exogenously):

$$\ln\left(\frac{Y_i}{L_i}\right) = \ln A_i = \ln \bar{A}_i + \alpha \ln L_i$$

- The agglomeration elasticity α is the object of much interest
- If we interpret GHM's effects as working this way then
  - By 5 years, local manuf. plants (0.27 share of local L) see TFP rise by 14%
  - Greenstone and Moretti (2004) use the same natural experiment and find (at 5 years) a 5% increase (though not very precisely estimated) in county-wide employment
  - So even if TFP effect is confined to non-MDP manuf. plants, could expect $\alpha \approx 0.7$
- But not at all obvious this is the right mechanism to have in mind

Further comments on GHM (2010)

- Policy:
  - Was it worth it for a typical county to (use subsidies to) attract an MDP?
  - Greenstone and Moretti (2004) use the same natural experiment to estimate the effect on total county land values. It's positive (and pretty big).
  - To the extent that all factors other than land are fully mobile, this should capitalize all of the local net benefits of the MDP.
  - (But could also include fiscal revenues from non-local taxes.)

- Subsequent findings:
  - Giroud et al (2022): not much spatial TFP spillover onto nearby plants, but large (same ballpark as main GHM effect) spillover within multi-regional firms who have a plant in the MDP county and some other county elsewhere (even far away)
  - Monte et al (AER, 2018): use the MDP shock to commuting elasticity (e.g. work in MDP county but don't live there)

Kline and Moretti (QJE, 2014)

- A different identification strategy for estimating agglomeration externalities: the 1933-onwards Tennessee Valley Authority
  - Enormous "place-based policy"
  - Perhaps one of the best examples of a "big push" policy ever tried
  - Famous episode in post-Depression (New Deal, FDR, etc) history
- KM (2014) uses this policy to generate quasi-experimental variation in local size, and hence to estimate agglomeration externalities. But, importantly, they also:
  - Of course estimate the direct effect of the TVA, which is of substantial independent interest
  - Also ask whether the agglomeration externalities take the form that is required for TVA to have an additional impact on national welfare through the fact that it promoted agglomeration.

What exactly was the TVA?

- A big (see Fig 1), ongoing (see Fig 2) place-based policy
- Key components:
  - Lots of public investment in infrastructure—hydroelectric dams, 650-mile navigation canal (1939-45), extensive road network (mostly done by 1950s), new schools, flood-control systems
  - Electricity sold inside TVA at reduced rates
- Which counties were selected into TVA?
  - See Table 1
  - Also potential additional "valley authorities" discussed in Congress in 1940s/1950s but never authorized. KM (2014) construct map of these regions (see Fig A2) based on their reading of the written proposals. Treat this as a placebo.

TVA region

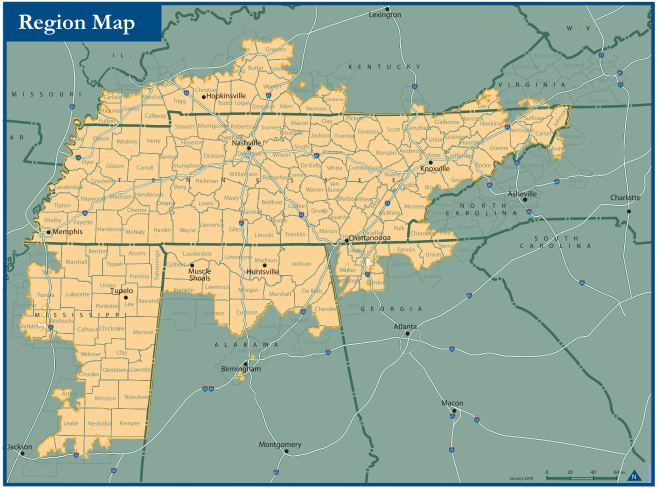


FIGURE I

The TVA Service Area (as of 2010)

TVA spending over time

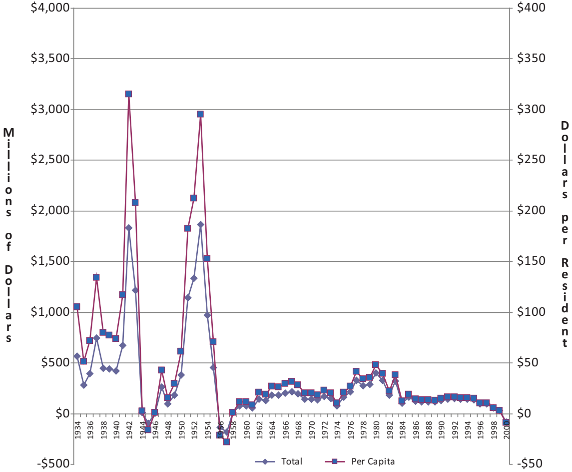


FIGURE II

Federal Transfers to TVA by Year (2000 Dollars)

Federal transfers defined as net federal outlays plus property transfers minus repayments (see Online Appendix for sources).

TVA Covariates (i.e. "Balance Table")

TABLE I

SUMMARY STATISTICS

	(1)	(2)	(3)	(4)	(5)	(6)
	Overall				Trimmed sample	
	TVA	Non-TVA	Non-TVA South	Non-TVA proposed authorities	Non-TVA	Non-TVA South
1930 characteristics						
Log population	9.991	9.977	9.989	9.940	9.905	9.979
Log employment	8.942	8.967	8.959	8.908	8.881	8.947
Log # of houses	8.445	8.508	8.455	8.466	8.442	8.445
Log average manufacturing wage	1.406	1.802	1.545	1.685	1.728	1.538
Manufacturing employment share	0.075	0.090	0.080	0.077	0.080	0.078
Agricultural employment share	0.617	0.455	0.541	0.510	0.487	0.547
% White	0.813	0.885	0.722	0.830	0.863	0.724
% Urbanized	0.153	0.280	0.233	0.216	0.242	0.215
% Illiterate	0.088	0.045	0.092	0.060	0.051	0.092
% of Whites foreign born	0.002	0.059	0.013	0.020	0.030	0.011
Log average farm value	5.252	5.646	5.386	5.552	5.579	5.370
Log median housing value	9.271	9.581	9.360	9.452	9.516	9.358
Log median contract rent	8.574	9.030	8.679	8.834	8.934	8.672
% Own radio	0.079	0.296	0.114	0.210	0.256	0.112
Max elevation (meters)	1,576.190	2,364.531	1,068.943	1,758.893	2,044.656	1,070.334
Elevation range (max–min)	1,127.761	1,521.322	712.336	1,083.293	1,251.074	715.253
% Counties in South	1.000	0.342	1.000	0.554	0.447	1.000

TVA Covariates (i.e. “Balance Table”)

TABLE I

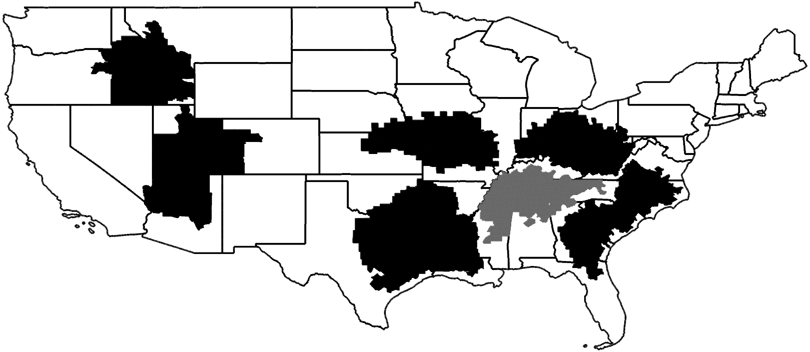
(CONTINUED)

	(1)	(2)	(3)	(4)	(5)	(6)
	Overall				Trimmed sample	
	TVA	Non-TVA	Non-TVA South	Non-TVA proposed authorities	Non-TVA	Non-TVA South
Changes 1920–1930						
Log population	0.051	0.049	0.067	0.004	0.037	0.060
Log employment	0.082	0.096	0.111	0.045	0.083	0.103
Log # of houses	0.078	0.092	0.108	0.046	0.078	0.100
Log average manufacturing wage	0.117	0.217	0.108	0.172	0.197	0.103
Manufacturing employment share	−0.010	−0.035	−0.018	−0.018	−0.026	−0.018
Agricultural employment share	−0.047	−0.036	−0.047	−0.046	−0.042	−0.047
% White	0.012	−0.011	−0.010	0.000	−0.006	−0.004
% Urbanized	0.047	0.064	0.080	0.042	0.054	0.069
% Illiterate	−0.030	−0.014	−0.029	−0.019	−0.015	−0.028
% of Whites foreign born	−0.001	−0.023	−0.016	−0.012	−0.015	−0.012
Log average farm value	−0.013	−0.076	0.025	−0.182	−0.102	0.013
# of Observations	163	2,326	795	828	1744	779
# of States	6	46	14	25	43	14

Notes. The unit of observation is a county. The trimmed sample is obtained by dropping control counties which, based on their preprogram characteristics, have a predicted probability of treatment in the bottom 25%. All monetary values are in constant 2000 dollars. Data are from the 1920 and 1930 Census of Population and Housing, with the exception of farm value data, which are from the 1920 and 1930 Agricultural Census, and elevation data, which were collected by Fishback, Haines, and Kantor (2007). Manufacturing wage is obtained by dividing the total annual wage bill in manufacturing by the estimated number of workers in the industry. Details on data construction and limitations are provided in the Online Appendix.

Alternative "Valley Authorities" (Placebo)

Figure A2: Map of Proposed Authorities



Notes: The map displays in black the six proposed authorities: the Atlantic Seaboard Authority, the Great Lakes-Ohio Valley Authority, the Missouri Valley Authority, the Arkansas Valley Authority, the Columbia Authority, and the Western Authority. The TVA region is displayed in gray.

Estimates of agglomeration externalities

- KM introduce a simple model of spatial equilibrium (a la Rosen-Roback; Roback, 1982). Ingredients:
  - Production function $Y_{it} = A_{it} K_{it}^{\alpha} F_i^{\beta} L_{it}^{1-\alpha-\beta}$, where capital K and labor L are assumed to be freely mobile and fixed factor F_i is not.
  - Output of this good is freely traded globally. But locations offer exogenous amenities $\bar{u}_{it}$ so labor mobility implies $\bar{u}_{it} w_{it}$ is equalized across all i within each t.
  - TFP given by:

$$\ln A_{it} = g\left(\frac{L_{it-1}}{R_i}\right) + \delta_t D_i + \eta_i + \gamma_t + \varepsilon_{it}$$

  - Where $g(\cdot)$ is the agglomeration function (assumed to be a function of *lagged* population density; R_i is area of county), and D_i is TVA treatment dummy
- So TVA has direct effects (that vary over time, δ_t) and potentially also indirect effects via agglomeration (i.e. via attracted L_{it-1} and $g'(\cdot) > 0$).

More structural estimation details

- Letting ε_{it} be a unit root (as Blanchard and Katz, 1992, found) with $\Delta\varepsilon_{it} = \lambda X_i + \nu_{it}$, can write SR labor demand function as (with "tilde" denoting original variables divided by β):

$$
\begin{aligned}
\ln L_{it} - \ln L_{it-1} &= -\frac{1-\alpha}{\beta}(\ln w_{it} - \ln w_{it-1}) + (\widetilde{\delta}_t - \widetilde{\delta}_{t-1})D_i \\
&+ \frac{1}{\beta}\left[g\left(\frac{L_{it-1}}{R_i}\right) - g\left(\frac{L_{it-2}}{R_i}\right)\right] + \widetilde{\lambda}X_i \\
&+ \widetilde{\gamma}_t - \widetilde{\gamma}_{t-1} + \widetilde{\nu}_{it}
\end{aligned}
$$

- In practice:
  - Proxy $g(x)$ with three-knot spline in $\ln(x)$
  - IV for each spline term k with lagged instrument: $Z_{it}^{(k)} \equiv g_k\left(\frac{L_{it-2}}{R_i}\right) - g_k\left(\frac{L_{it-3}}{R_i}\right)$
  - Calibrate (SR) LD elasticity $-\frac{1-\alpha}{\beta}$ from labor literature (Hammermesh 1993 chapter, values: 1-1.5)

KM (2014): Estimates of $\frac{1}{\beta}g(\cdot)$ (and $\beta = 0.47$)

TABLE VI

STRUCTURAL ESTIMATES OF AGGLOMERATION FUNCTION (LOG BASIS)

	(1) OLS	(2) OLS	(3) OLS	(4) 2SLS	(5) 2SLS	(6) 2SLS
Change in log manufacturing density spline components:						
Low	0.173	0.147	0.146	0.443	0.400	0.396
	(0.037)	(0.037)	(0.037)	(0.102)	(0.108)	(0.107)
				[177.17]	[159.14]	[157.20]
Medium	0.221	0.227	0.226	0.456	0.440	0.438
	(0.045)	(0.044)	(0.045)	(0.124)	(0.123)	(0.124)
				[106.74]	[109.55]	[110.13]
High	0.143	0.012	0.141	0.466	0.467	0.453
	(0.051)	(0.050)	(0.050)	(0.150)	(0.150)	(0.151)
				[206.66]	[204.69]	[200.36]
Log manufacturing wages	−1.5	−1.5	−1.5	−1.5	−1.5	−1.5
TVA	0.007	0.012	0.008	−0.003	0.002	−0.002
	(0.014)	(0.014)	(0.014)	(0.012)	(0.013)	(0.012)
Regional trends	no	no	yes	no	no	yes
1940 manufacturing density	no	yes	yes	no	yes	yes
Decade effects	yes	yes	yes	yes	yes	yes
Controls for 1920 and 1930 characteristics	yes	yes	yes	yes	yes	yes
p-value equal slopes	0.2483	0.1298	0.1038	0.9545	0.6695	0.7171
p-value slopes equal 0	0.0000	0.0000	0.0000	0.0002	0.0007	0.0016
N	6,057	6,057	6,057	5,935	5,935	5,935

Notes. Dependent variable is change in log county manufacturing employment. Manufacturing density is manufacturing employment per square mile. Standard errors clustered by state in parentheses. Angrist-Pischke cluster robust first stage F-statistic in brackets. All estimates weighted by 1950 county population. "Low" refers to spline component corresponding to log density below 60th percentile of 1980 distribution, "Medium" to log density between 60th and 85th percentile of 1980 distribution, and "High" to log density above 85th percentile of 1980 distribution. Spline coefficients give the elasticity of labor demand with respect to lagged manufacturing density over the relevant range. The instruments are changes in the spline components of log manufacturing density lagged by two decades.

Peters (2022)

- Examines the post-WWII forced resettlement to West Germany of ethnic Germans living in Eastern Europe
  - Mostly 1945-48, but by 1950 8 million were resettled (20% of W. German pop)
  - “One of largest forced population movements in world history”
- Did this raise Y/L in areas of West Germany that saw inflows of refugees?

Peters (2022): The Setting

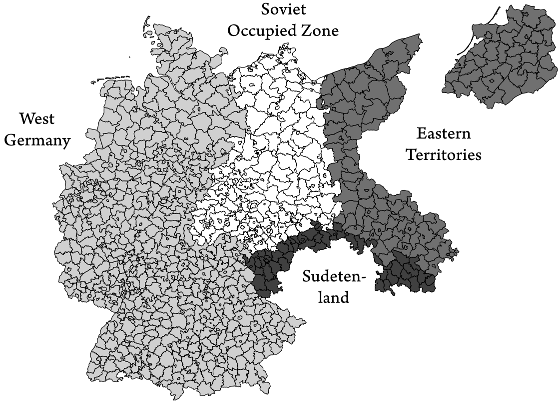


FIGURE 1.—The German Reich in 1939. *Note*: The figure shows the German Reich in the boundaries of 1939. The light-gray shaded part in the west is the area of to-be West Germany. The medium-gray shaded parts in the east are the Eastern Territories of the German Reich. The dark shaded area in the southeast is the Sudetenland. The white shaded part is the area of the Soviet Occupied Zone. The intra-regional spatial units are counties.

Peters (2022): Regional variation within West Germany

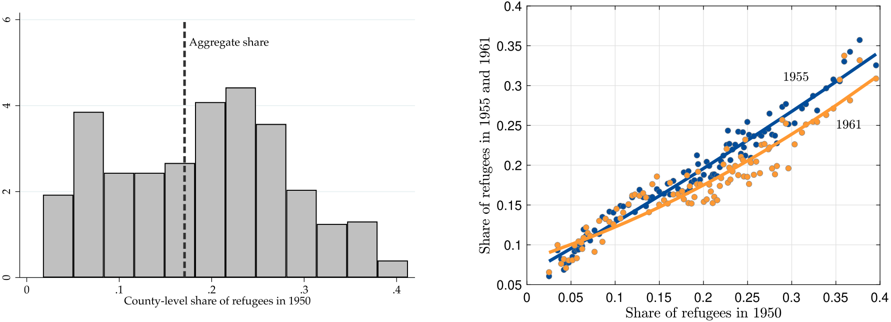


FIGURE 2.—The Heterogeneity and Persistence of Refugee Inflows. *Note*: The left panel shows the distribution of the share of refugees in 1950 across counties. The right panel shows the correlation between the share of refugees in 1950 and 1955 (blue) and 1961 (orange) as binned scatter plots for 100 percentiles of the refugee share in 1950.

Peters (2022): Sources of pop. changes in W. Germany

TABLE III

THE POPULATION OF WEST GERMANY: 1939–1950.

	Population Losses 1939–1950				Population Gains 1939–1950			
Population 1939	Military Losses	Civilian Losses	Non-military Deaths	Others	Refugees	Inflows from SOZ	Births	Population 1950
39.3 m	2 m	0.4 m	5.2 m	0.5 m	7.9 m	1.5 m	7 m	47.6 m

Note: The table reports aggregate population trends in West Germany between 1939 and 1950. "Inflows from SOZ" are individuals who fled the Soviet Occupied Zone. Source: Edding (1951, p. 2).

Peters (2022): Characteristics of refugees

TABLE IV

CHARACTERISTICS OF REFUGEES AND NATIVES.

	Male Share	Age Distribution			Educational Attainment			
		<15	20–65	65+	Elem. School	High School	Vocational School	College
Natives	46.5	20.4	68.5	11.1	66.8	26.3	4.9	1.9
Refugees	46.9	21.9	68.5	9.7	67.4	25.6	4.9	2.1

Note: The first panels report the share of males and the age distribution in 1958. The last panel reports the distribution of educational attainment of the cohort born before 1920 as observed in the 1970 census. These individuals were at least 25 years old in 1945 and hence completed their educational attainment prior to the expulsion. Source: Besser (2007).

Peters (2022): Effects on Y/L and L

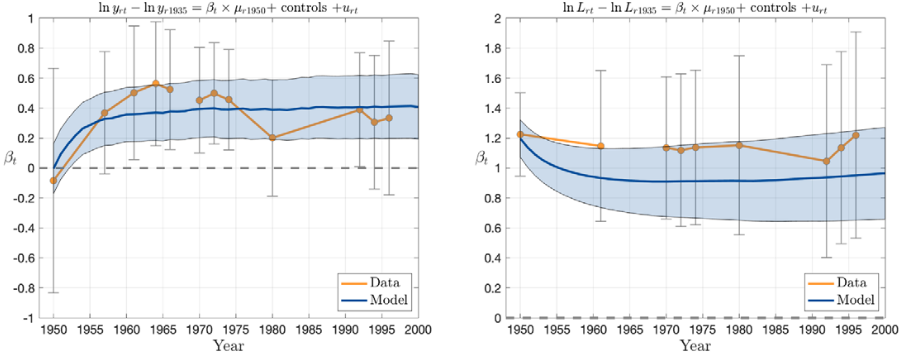


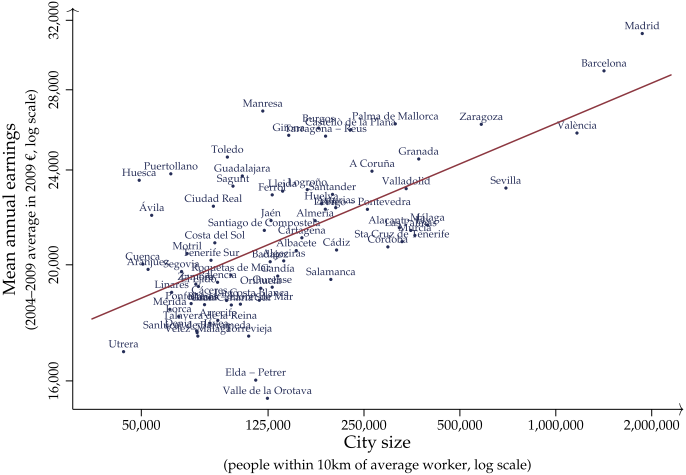
FIGURE 4.—Model Fit: the Dynamic Effects on Income and Population Growth. *Note*: The figures report the coefficient β of the regression $y_{rt} = \beta_t \mu_{r1950} + x'_{rt}\gamma + u_{rt}$ for different time horizons and for income growth (left panel) and population growth (right panel) as dependent variables. The vector x_{rt} controls for state fixed effects, population density in 1939, war-time destruction, log income per capital (population) in 1939, and the distance to the inner German border (see columns 2 or 6 of Table VI). For both the model and the data, I also report 95% confidence intervals.

Complementary evidence from "mover designs"

- The research designs we have studied so far have involved changing the attributes of a "place" and then observing the impact on economic outcomes
- An alternative is to track how individual-level outcomes (e.g. wages) change as individuals change their place
- This is the strategy pursued in several recent papers
    - Glaeser and Mare (JoLE, 2001)
    - Combes, Duranton and Gobillon (JUE, 2018)
    - de la Roca and Puga (REStud, 2017)
    - Card, Rothstein and Yi (2021)
- We will focus on the last two (which extend ideas in the earlier studies)

de la Roca and Puga (2017): Nominal wages are higher (on average) in big cities

(Also replicated many times in many countries; see recitation for Glaeser and Gottlieb's (JEL 2009) treatment for US cities.)



Mover designs

- The basic idea (following the influential analog for estimating firm-specific effects on individual-level wages, in Abowd, Kramarz and Margolis (ECMA, 1999)) is to estimate

$$\ln w_{ict} = \alpha_c + \mu_i + \beta x_{it} + \varepsilon_{ict}$$

- Where
    - w_{ict} is the *nominal* wage of individual i who works in city c at time t
    - α_c is a city-level fixed effect
    - μ_i is an individual-level fixed effect
    - x_{it} is a vector of time-varying *observed* components of worker "skill" (age, experience, perhaps education if we can observe it, etc.)

Mover designs

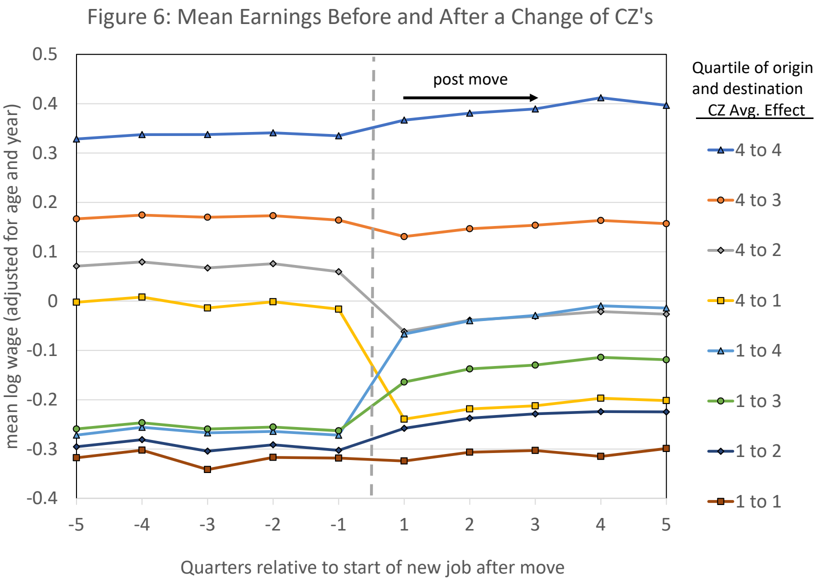
- The basic idea (following the influential analog for estimating firm-specific effects on individual-level wages, in Abowd, Kramarz and Margolis (ECMA, 1999)) is to estimate

$$\ln w_{ict} = \alpha_c + \mu_i + \beta x_{it} + \varepsilon_{ict}$$

- Where
  - w_{ict} is the *nominal* wage of individual i who works in city c at time t
  - α_c is a city-level fixed effect
  - μ_i is an individual-level fixed effect
  - x_{it} is a vector of time-varying *observed* components of worker "skill" (age, experience, perhaps education if we can observe it, etc.)

- If our goal is to identify α_c (e.g. to then project these on city sizes), which individuals contribute to its identification?
- Under what assumptions could we hope to obtain consistent (for fixed C, but large N and/or T) estimates of α_c?

Card et al (2021): “Event Studies” figure



de la Roca and Puga (2017): "Event Study" figure

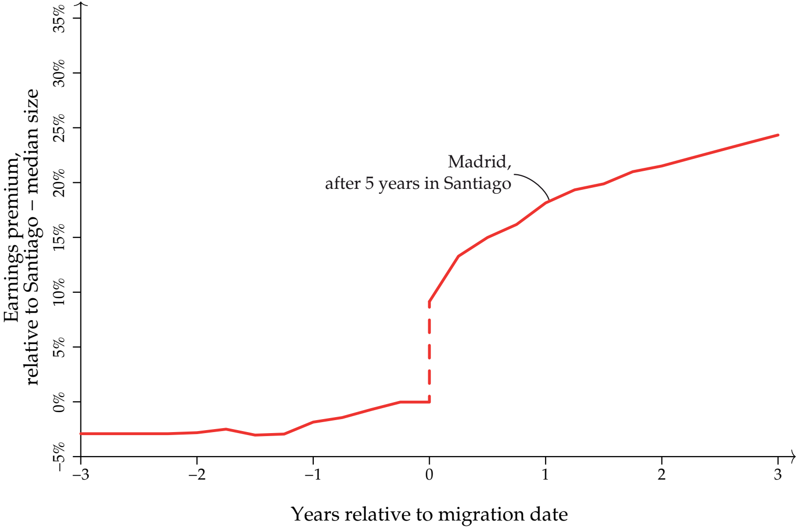
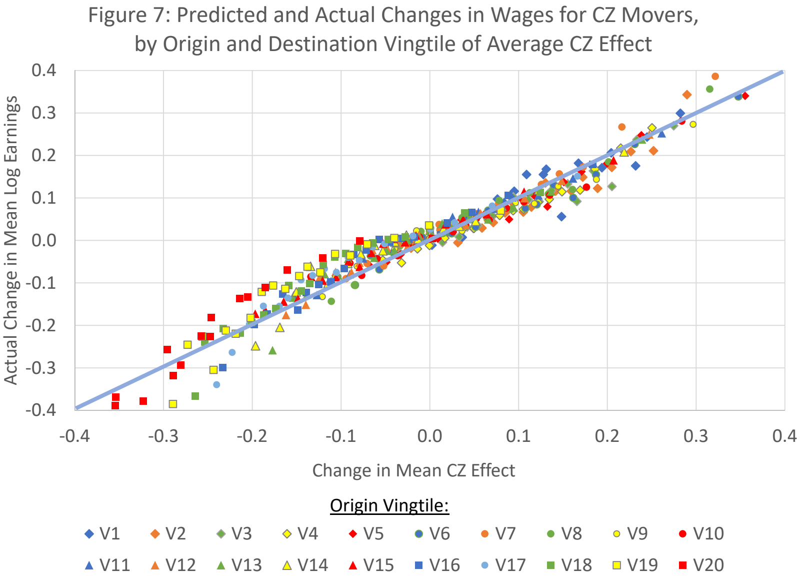


FIGURE 6

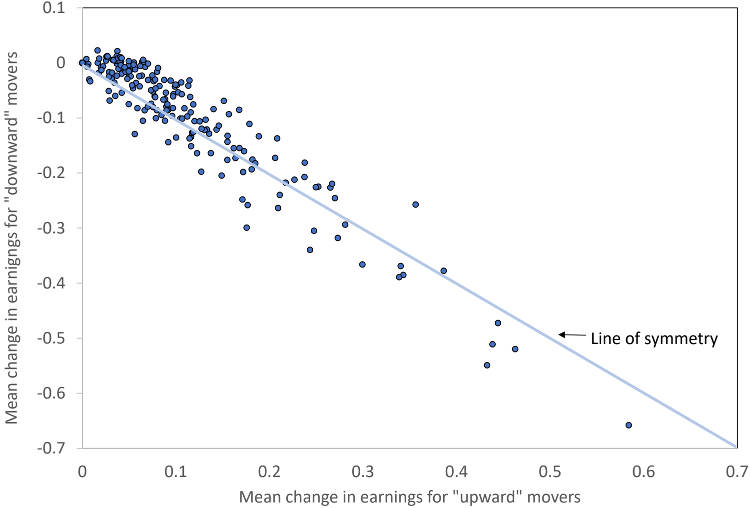
Non-parametric pre- and post-migration earnings profile relative to median-sized city

Card et al (2021): Test of AKM-like functional form



Card et al (2021): Test for symmetry of AKM-like functional form

Figure 8: Evaluation of Symmetry for CZ changers



de la Roca and Puga (2017): Controlling for individual-level heterogeneity

TABLE 1
Estimation of the static city size earnings premium

	(1)	(2)	(3)	(4)
Dependent variable	Log earnings	City indicator coefficients column (1)	Log earnings	City indicator coefficients column (3)
Log city size		0.0455 (0.0080)***		0.0241 (0.0058)***
City indicators	Yes		Yes	
Worker fixed effects	No		Yes	
Experience	0.0319 (0.0005)***		0.1072 (0.0018)***	
Experience^2	−0.0006 (0.0000)***		−0.0014 (0.0000)***	
Firm tenure	0.0147 (0.0006)***		0.0042 (0.0004)***	
Firm tenure^2	−0.0005 (0.0000)***		−0.0003 (0.0000)***	
Very-high-skilled occupation	0.7752 (0.0062)***		0.2350 (0.0057)***	
High-skilled occupation	0.4976 (0.0046)***		0.1758 (0.0040)***	
Medium-high-skilled occupation	0.2261 (0.0031)***		0.0873 (0.0029)***	
Medium-low-skilled occupation	0.0542 (0.0021)***		0.0152 (0.0019)***	
University education	0.2014 (0.0037)***			
Secondary education	0.1084 (0.0022)***			
Observations	6,263,446	76	6,263,446	76
R^2	0.4927	0.2406	0.1144	0.1422

(This drop by about 50% is common in other settings too.)

Extending the mover design

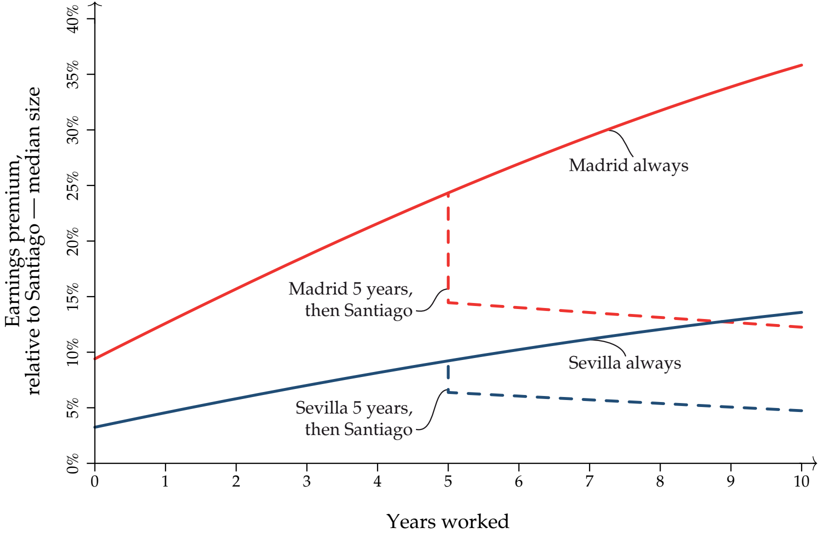
- de la Roca and Puga (2017) actually focus most on a dynamic specification:

$$\ln w_{ict} = \alpha_c + \mu_i + \sum_j \delta_{jc} e_{ijt} + \beta x_{it} + \varepsilon_{ict}$$

- Where
  - e_{ijt} is the number of years that individual i has spent working in city j up to date t
  - δ_{jc} is a city pair-specific return-to-experience parameter (how much wages when working in c are affected by prior experience in any city j

- Why do this?

de la Roca and Puga (2017): “Learning by working in big cities”



Panel (a) Profiles allowing for learning benefits of bigger cities

de la Roca and Puga (2017)

TABLE 2
Estimation of the dynamic and static city size earnings premia

	(1)	(2)	(3)
Dependent variable	Log earnings	Initial premium (city indicator coefficients column (1))	Medium-term premium (initial + 7.7 years local experience)
Log city size		0.0223 (0.0058)***	0.0510 (0.0109)***
City indicators	Yes		
Worker fixed effects	Yes		

(So the 50% drop we saw earlier wasn't due to sorting, but due to omitting the city-specific experience effects.)

de la Roca and Puga (2017): Surprisingly little sorting

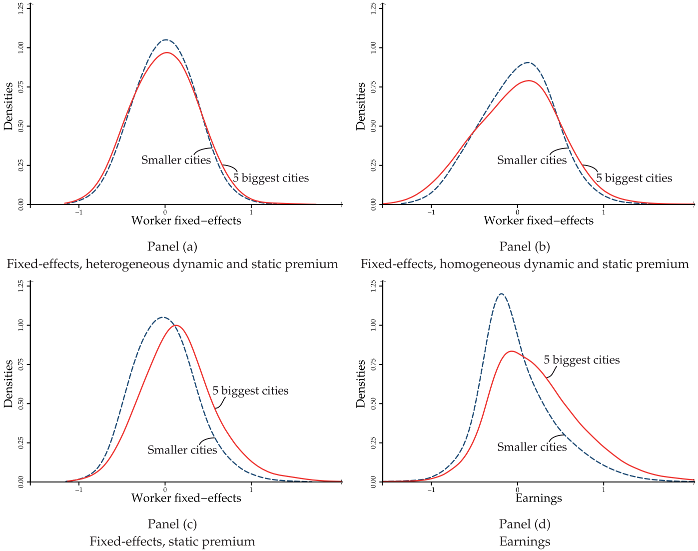


FIGURE 8

Comparisons of worker fixed-effects distributions across cities

de la Roca and Puga (2017): Do workers learn (from big cities) at heterogeneous rates?

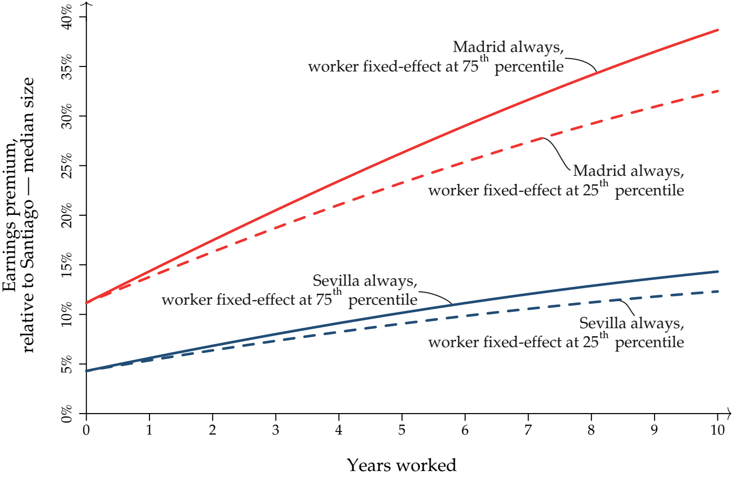


FIGURE 7

Earnings profiles relative to median-sized city, high- and low-ability worker

More on Card et al (2021)

- CRY extend the de la Roca and Puga (2017) model to model the "city" (c) effect, α_c, to actually be a city-times-industry effect, α_{ck}
- They then use cross-industry-and-city movers to test (and find numerous pieces of evidence consistent with) a model

$$\alpha_{ck} = \alpha_c + \alpha_k + \epsilon_{ck}$$

  where ϵ_{ck} is as good as random
- In other words: NYC pays more than NOLA, not because it has more workers in finance (an industry with a relatively high wage nationally), but because workers in all of its industries (not just finance) are on average paid more than they are in the same industry in NOLA
- However, relative to de la Roca and Puga (2017), CRY do not estimate city-specific learning functions
  - So not clear whether these are rejected by the event study figures in CRY (i.e. learning by working in big cities more important in Spain than in US) or just not studied in CRY

Other Literature

- Firm-level heterogeneity across cities:
  - Combes, Duranton and Gobillon (ECMA, 2012)
  - Gaubert (AER, 2017)
- Consumption-side externalities:
  - Diamond (AER, 2016)
  - Handbury and Weinstein (REStud, 2015)

Ideas for Further Research

- Still need more/better natural experiments for estimating agglomeration functions
- Endogenous amenities/social spillovers have received much less attention than endogenous productivity
- Extensions of agglomeration functions that need more exploration:
  - More tests of the functional form beyond KM
  - Industry-specific functions
  - Factor-specific functions (E.g. skill-biased agglomeration effects? See Baum-Snow, Pavan and Freedman, 2018)
  - Different spatial scales
- Comparing micro- (e.g. worker-level) and macro- approaches (see Combes and Gobillon, 2015 survey)
- Is (all of) what we're identifying really an externality?
- Which mechanisms are more important for driving agglomeration externalities?
- Dynamics of agglomeration functions (e.g. Allen and Donaldson, 2019)