Spatial empirics

Yanos Zylberberg (Bristol, CEPR)

UEA Summer School, May, 2026

The course

This module will (superficially) cover the issue of spatial empirics.

1. Spatial modeling: combining structure and estimation?
   a. a stylized framework;
   b. the old issue of demand and supply estimation;
   c. the estimation of elasticities;
   d. the identification of externalities.

2. Spatial identification: a few approaches to spatial identification.
   a. difference-in-differences, SUTVA and general equilibrium;
   b. spatial discontinuity;
   c. diffusion processes;
   d. inference.

Spatial modeling

1.a. A stylized framework

A stylized framework

A researcher is interested in understanding the allocation of individuals across cities $\ell \in \mathcal{L}$.

This section presents the foundations of a quantitative spatial model designed to study:

- demand and supply estimation (and equilibrium variables);
- the estimation of elasticities;
- the identification of externalities (and exogenous fundamentals).

Location choice

Consider a unit mass of workers with idiosyncratic preferences over cities $\ell \in \mathcal{L}$.

A worker ψ residing in city ℓ obtains the following indirect utility:

$$u_\ell(\psi) = a_\ell(\psi) \frac{w_\ell}{q_\ell}$$

where $a_\ell(\psi)$ is an idiosyncratic preference, drawn independently across individuals and locations from a distribution $F_\ell(\cdot)$—parametrized by the city-specific amenity b_ℓ.

In short, workers value: (i) wages w_ℓ, (ii) local prices q_ℓ, (iii) the city-specific amenity b_ℓ, and (iv) their own idiosyncratic preferences.

How do workers choose among locations?

How do workers choose among locations?

Worker ψ chooses to live in the city that maximizes:

$$u_\ell(\psi) = a_\ell(\psi)\frac{w_\ell}{q_\ell}$$

For example, worker ψ prefers city ℓ to city ℓ' if,

$$a_\ell(\psi)\frac{w_\ell}{q_\ell} > a_{\ell'}(\psi)\frac{w_{\ell'}}{q_{\ell'}}$$

The presence of idiosyncratic preferences implies that (some) people are willing to live in any places, despite substantial differences in overall living conditions.

How do workers choose among locations?

Worker ψ chooses to live in the city that maximizes:

$$u_\ell(\psi) = a_\ell(\psi)\frac{w_\ell}{q_\ell}$$

For example, worker ψ prefers city ℓ to city ℓ' if,

$$a_\ell(\psi)\frac{w_\ell}{q_\ell} > a_{\ell'}(\psi)\frac{w_{\ell'}}{q_{\ell'}}$$

The presence of idiosyncratic preferences implies that (some) people are willing to live in any places, despite substantial differences in overall living conditions.

Notes: Idiosyncratic preferences may also be interpreted as a stylized representation of relocation frictions.

Most models use **extreme value** distributions for convenience (Turner).

With a **Gumbel** distribution (type 1, "multinomial logit"), we have that $F_\ell(a) = e^{-e^{-a+b_\ell}}$, and:

$$\pi_\ell = \frac{\exp(w_\ell + b_\ell - q_\ell)}{\sum_{\ell' \in \mathcal{L}} \exp(w_{\ell'} + b_{\ell'} - q_{\ell'})}$$

is the probability to prefer city ℓ.

With a **Fréchet** distribution (type 2):

- $F_\ell(a) = e^{-(a/b_\ell)^{-\chi}}$,
- $\chi > 1$, and b_ℓ captures the amenity in location ℓ.

$$\pi_\ell = \frac{(b_\ell w_\ell / q_\ell)^\chi}{\sum_{\ell' \in \mathcal{L}} (b_{\ell'} w_{\ell'} / q_{\ell'})^\chi}$$

1.b. The old issue of demand and supply estimation

The previous location-choice model implies that, given $\sum_{\ell' \in \mathcal{L}} l_{\ell'} = 1$:

$$\boxed{\ln(l_\ell) = \chi \ln(b_\ell) + \chi \ln(w_\ell / q_\ell) - \gamma} \quad (S)$$

where γ is an equilibrium term combining living conditions across all locations.

The previous location-choice model implies that, given $\sum_{\ell' \in \mathcal{L}} l_{\ell'} = 1$:

$$\boxed{\ln(l_\ell) = \chi \ln(b_\ell) + \chi \ln(w_\ell / q_\ell) - \gamma} \quad (S)$$

where γ is an equilibrium term combining living conditions across all locations.

Suppose that the researcher instead estimates:

$$\boxed{\ln l_\ell = \alpha + \beta \ln b_\ell + \varepsilon_\ell}$$

where l_ℓ denotes the population of city ℓ and b_ℓ is a measure of local amenities, with the goal of interpreting β as an estimate of χ.

The previous location-choice model implies that, given $\sum_{\ell' \in \mathcal{L}} l_{\ell'} = 1$:

$$\boxed{\ln(l_\ell) = \chi \ln(b_\ell) + \chi \ln(w_\ell / q_\ell) - \gamma} \quad (S)$$

where γ is an equilibrium term combining living conditions across all locations.

Suppose that the researcher instead estimates:

$$\boxed{\ln l_\ell = \alpha + \beta \ln b_\ell + \varepsilon_\ell}$$

where l_ℓ denotes the population of city ℓ and b_ℓ is a measure of local amenities, with the goal of interpreting β as an estimate of χ.

What is the **fundamental problem** in this transformation of the previous structural equation?

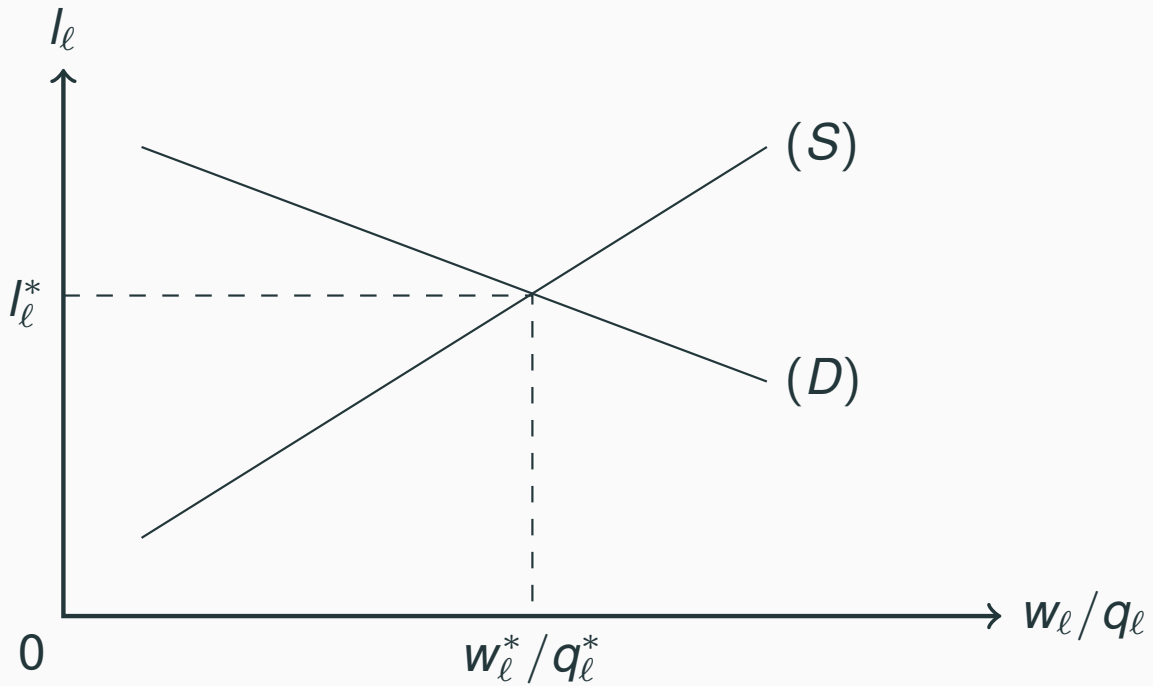
The key issue is omitted variation in estimating the **labor supply** equation, while ignoring market equilibrium.

The key issue is omitted variation in estimating the **labor supply** equation, while ignoring market equilibrium. Indeed, assume that the tradable good (numeraire) in location ℓ is produced following $y_\ell = \mathcal{T}_\ell \cdot l_\ell^{1-\alpha}$:

$$\ln w_\ell = \ln \mathcal{T}_\ell + \ln(1-\alpha) - \alpha \ln l_\ell$$

is the (inverse) **labor demand** resulting from profit maximization. Housing production would lead to a similar equation: $\ln q_\ell = \eta + \sigma \ln l_\ell$.

$$\boxed{\ln w_\ell / q_\ell = \ln \mathcal{T}_\ell - \eta + \ln(1-\alpha) - (\alpha+\sigma) \ln l_\ell} \quad (D)$$



The key issue is omitted variation in estimating the **labor supply** equation, while ignoring market equilibrium. Indeed, assume that the tradable good (numeraire) in location ℓ is produced following $y_\ell = \mathcal{T}_\ell \cdot l_\ell^{1-\alpha}$:

$$\ln w_\ell = \ln \mathcal{T}_\ell + \ln(1-\alpha) - \alpha \ln l_\ell$$

is the (inverse) **labor demand** resulting from profit maximization. Housing production would lead to a similar equation: $\ln q_\ell = \eta + \sigma \ln l_\ell$.

$$\boxed{\ln w_\ell / q_\ell = \ln \mathcal{T}_\ell - \eta + \ln(1-\alpha) - (\alpha+\sigma) \ln l_\ell} \quad (D)$$

Combining **labor demand** and **labor supply**:

$$\ln l_\ell = \alpha + \beta \ln b_\ell + \beta \ln \mathcal{T}_\ell$$

where:

$$\beta = \frac{\chi}{1+\chi(\alpha+\sigma)}$$

$$\alpha = \frac{\chi(\ln(1-\alpha)-\eta)}{1+\chi(\alpha+\sigma)}$$

The key parameter in understanding the responsiveness of individuals to local conditions is χ, but estimating the following relationship,

$$\boxed{\ln l_\ell = \alpha + \beta \ln b_\ell + \varepsilon_\ell}$$

to **uncover β as an estimate for χ** is not the right approach.

Equilibrium effects across geographies induces an omitted variable bias in the previous estimation.

How can we estimate such elasticity then?

1.c. The estimation of “elasticities”

The general idea is that the **structure of a model** can help in the estimation of key elasticities.

In the previous example, location choice follows:

$$\boxed{\ln(l_\ell) = \chi \ln(b_\ell) + \chi \ln(w_\ell / q_\ell) - \gamma}$$

The best approach that exploits the model structure is to:

i. control for observable amenities b_ℓ,

ii. and exploit variation in w_ℓ / q_ℓ that is exogenous to labor supply considerations.

What would be such variation?

In Imbert et al. [2023], we estimate the probability to migrate across Chinese cities $\ell \in \mathcal{L}$ from rural origins $o \in \mathcal{O}$:

$$\ln \pi_{o\ell} = \delta_o + \chi \ln \underbrace{w_\ell / q_\ell}_{v_{o\ell}} + \beta \mathbf{X}_{o\ell} + \varepsilon_{o\ell} \quad (S)$$

where real wages across destinations are instrumented with a trade-induced **labor demand** shifter,

$$\ln w_\ell / q_\ell = d_o + a T_\ell + b \mathbf{X}_{o\ell} + e_{o\ell} \quad (D)$$

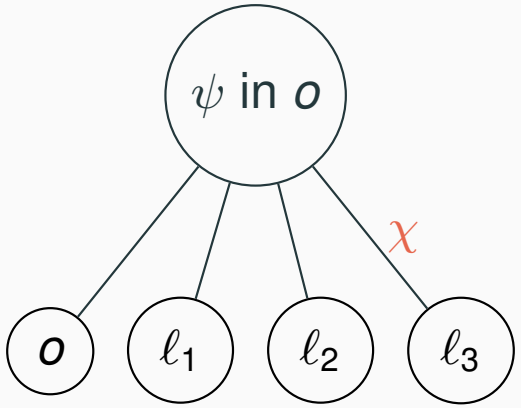
intuitively capturing exogenous fluctuations in $\ln \mathcal{T}_\ell$.

	Father alone (1)	Mother alone (2)	Both parents (3)	Whole family (4)	Others (5)
Real wage	14.093	16.741	10.145	6.424	11.143
	(2.654)	(2.585)	(1.611)	(1.229)	(2.168)
Observations	74,031	74,031	74,031	74,031	74,031
Migration mode	$j = 1$	$j = 2$	$j = 3$	$j = 4$	o
F-stat	19.41	25.60	18.83	15.72	17.98

Notes: The unit of observation is an origin-destination prefecture pair in 2005. The specification is estimated using a two-stage Poisson pseudo-maximum likelihood (PPML-IV) estimator, with population weights at origin in 2000. Standard errors (reported in parentheses) are two-way clustered at the origin and destination levels. The explanatory variable is the value at destination, instrumented using a trade-induced labor demand shock. Controls include origin fixed effects, (log) population at destination in 2000, and (log) distance between origin and destination.

A 1% increase in real wages increases migrant flows to a location by 6–14%.

With the previous elasticities and a simple logit structure...

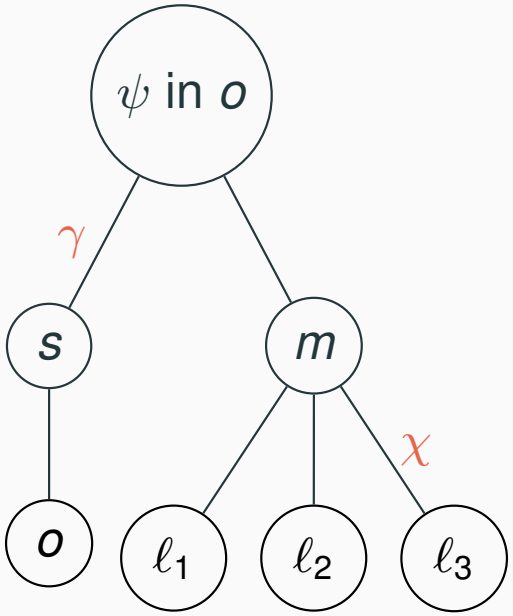


Notes: The Figure represents the structure induced by our assumptions on the distribution of idiosyncratic preferences for relocation, $\varepsilon_{\mathbf{k}}$. The parameter χ is the shape parameter for the extreme value distribution.

...the doubling of real wages across Chinese cities between 2000–2005 would have led to the emptying of all rural areas.

What are we missing?

The previous model $\pi_{o\ell} = \left(\frac{V_\ell}{V}\right)^\chi$, $V = \left[V_o^\chi + \sum_{\ell\in\mathcal{L}} (V_\ell)^\chi\right]^{1/\chi}$ does not work. We can assume instead that preferences follow a **nested tree**:



Notes: The Figure represents the nested structure induced by our assumptions on the distribution of idiosyncratic preferences for relocation, $\varepsilon_{\mathbf{k}}$. The parameter γ is the shape parameter for the upper nest of the generalized extreme value distribution; and the parameter χ is the shape parameter for the lower nest of the generalized extreme value distribution.

where the probability $\pi^m_{o\ell}$ behaves like a Frechet distribution of parameter χ:

$$\pi_{o\ell} = \underbrace{\left(\frac{V_o^m}{V_o}\right)^\gamma}_{\pi_o^m} \underbrace{\left(\frac{V_{o\ell}^m}{V_o^m}\right)^\chi}_{\pi^m_{o\ell}}, \quad V_o^m = \left[\sum_{\ell\in\mathcal{L}} (V_{o\ell}^m)^\chi\right]^{1/\chi}, \quad V_o = \left[(V_o^m)^\gamma + (V_o^s)^\gamma\right]^{1/\gamma}$$

In the previous model, one can estimate the probability to migrate across Chinese cities—**conditional on migrating**:

$$\ln \pi^m_{o\ell} = \delta_o + \chi \ln w_\ell / q_\ell + \beta \mathbf{X}_{o\ell} + \varepsilon_{o\ell} \quad (S)$$

as before.

The estimation of the upper nest (γ) of the location model relies on the following structural equation:

$$\ln \left(\frac{\pi^m_o}{\pi^s_o} \right) = \gamma \ln \left(\frac{V^m_o}{V^s_o} \right)$$

In the previous model, one can estimate the probability to migrate across Chinese cities—**conditional on migrating**:

$$\ln \pi^m_{o\ell} = \delta_o + \chi \ln w_\ell / q_\ell + \beta \mathbf{X}_{o\ell} + \varepsilon_{o\ell} \quad (S)$$

as before.

The estimation of the upper nest (γ) of the location model relies on the following structural equation:

$$\ln \left(\frac{\pi^m_o}{\pi^s_o} \right) = \gamma \ln \left(\frac{V^m_o}{V^s_o} \right)$$

How would you identify γ?

We consider the following empirical counterpart:

$$\boxed{\ln\left(\frac{\pi_o^m}{\pi_o^s}\right) = a + b\ln\left(\frac{V_o^m}{V_o^s}\right) + \varepsilon_o} \quad (S)$$

where $V_o^s = w_o/q_o$ and $V_o^m = \left[\sum_{\ell\in\mathcal{L}}(V_{o\ell}^m)^{\widehat{\chi}}\right]^{1/\widehat{\chi}}$. One can instrument $\ln\left(\frac{V_o^m}{V_o^s}\right)$ by an origin **labor demand** shifter!

$$\ln\left(\frac{V_o^m}{V_o^s}\right) = \alpha + \beta\omega_o + \varepsilon_o \quad (D)$$

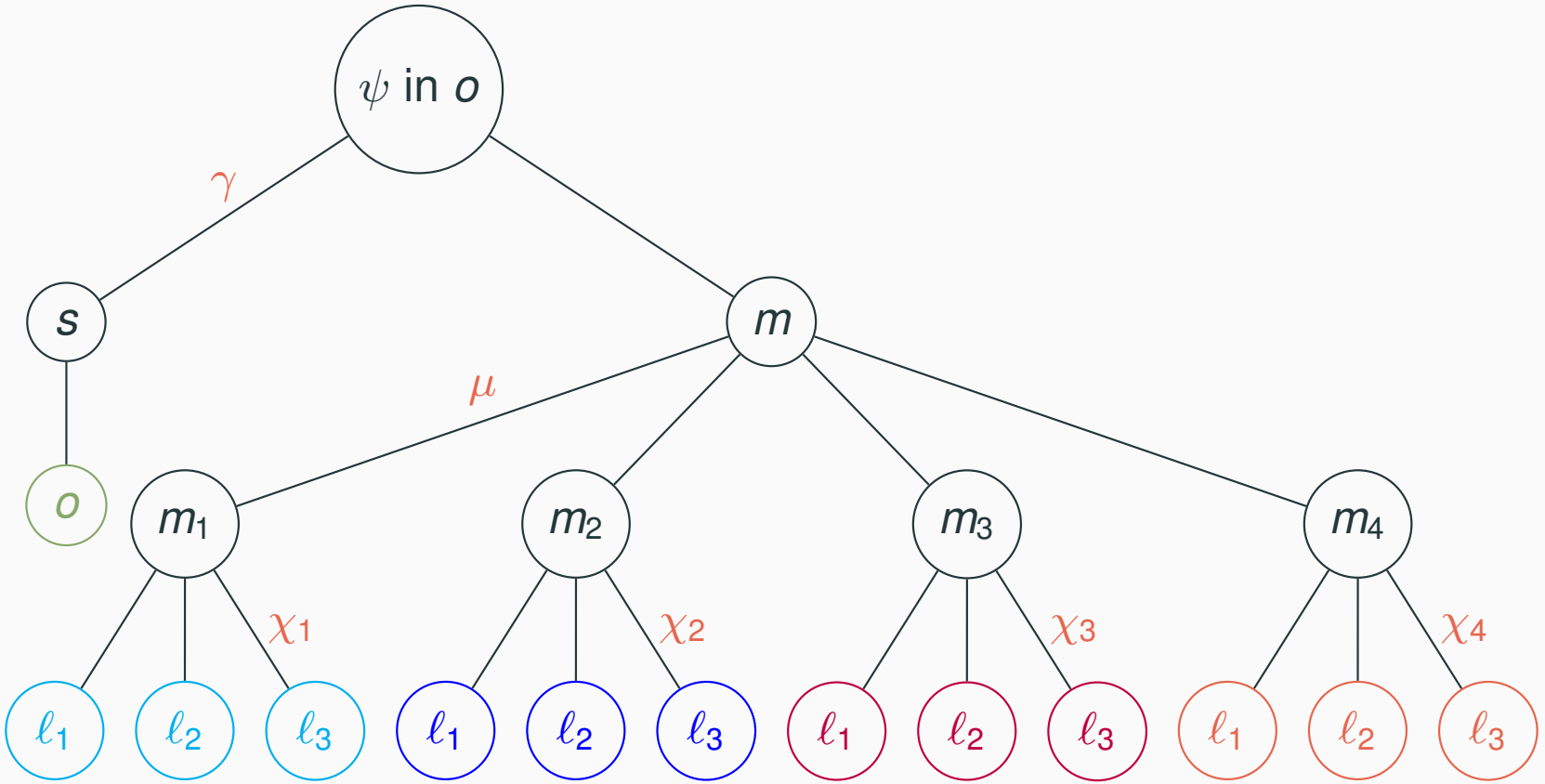
Parameters	χ	γ
	11.143	2.371
	(2.168)	(0.596)
	[17.98]	[15.03]
	74,031	257

Notes: This table reports estimates for the location choice model. The first column reports estimates from the lower nest (χ). The second column reports the estimate from the upper nest (γ).

A 1% increase in urban real wages increases rural migrant flows by 2.4%.

In fact, Imbert et al. [2023] formulate a nested choice model capturing whether ($m \in \{0, 1\}$), how ($j \in \{1, 2, 3, 4\}$), and where (ℓ) households relocate:

$$\max_{\mathbf{k}} \{\ln V_{\mathbf{k}} + \varepsilon_{\mathbf{k}}\},$$



Notes: The parameter γ is the shape parameter for the upper nest of the generalized extreme value distribution; the parameter μ is the shape parameter for the intermediate nest of the generalized extreme value distribution; and the parameters (χ_j) are the shape parameters for the lower nests of the generalized extreme value distribution.

Question from the audience: “Respectable sir, why is it important to estimate credible elasticities rather than calibrate them using estimates from previous studies?”

Question from the audience: “Respectable sir, why is it important to estimate credible elasticities rather than calibrate them using estimates from previous studies?”

a. Internal validity:

- cross-location elasticities estimated in other settings may differ substantially;
- counterfactual predictions may therefore be quantitatively misleading.

Question from the audience: “Respectable sir, why is it important to estimate credible elasticities rather than calibrate them using estimates from previous studies?”

a. Internal validity:

- cross-location elasticities estimated in other settings may differ substantially;
- counterfactual predictions may therefore be quantitatively misleading.

b. Biases may spill over to other model parameters (partly because of the inversion approach):

- quantitative models are often exactly identified, in the sense that any discrepancy between the model and the data is absorbed by residual components (typically amenities or productivities);
- the model can mechanically fit the observed allocation even when key elasticities are misspecified;
- if the elasticity is set too low, there will be too few workers in cities that have high production potential; residual components will need to “explain” the actual allocation of workers skewed toward these places;
- for example, amenities would need to be artificially high in those cities.

1.d. The identification of externalities

Our previous exercise considered:

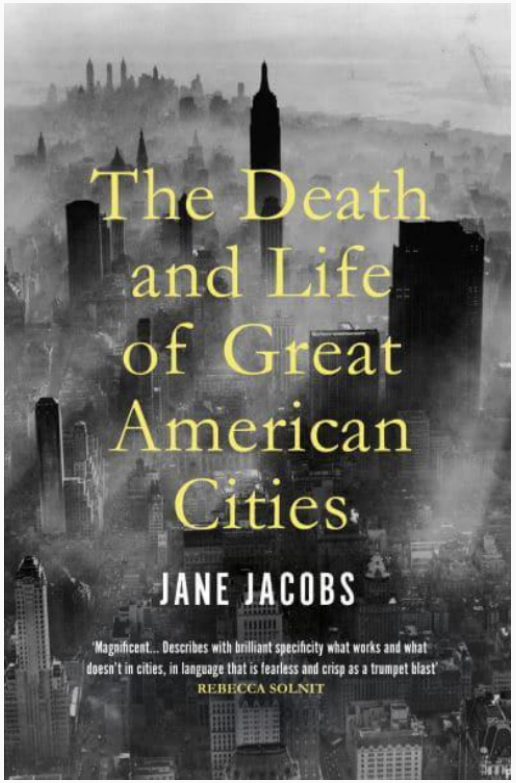
$$\boxed{\ln(l_\ell) = \chi \ln(b_\ell) + \chi \ln(w_\ell/q_\ell) - \gamma}, \quad \ln w_\ell/q_\ell = \ln \mathcal{T}_\ell + \zeta - (\alpha + \sigma) \ln l_\ell$$

focusing on χ, instead of amenities b_ℓ or productivities $\mathcal{T}_\ell$. However, urban economics has long been interested in:

- agglomeration spillovers (passing through fundamentals $\mathcal{T}_\ell$),
- congestion externalities (b_ℓ),
- industry-based externalities (Jacobs, MAR, $\mathcal{T}_\ell$),
- etc.

Conveniently, most quantitative models have the following **inversion property**: conditional on observing prices and/or quantities and elasticities, one can recover productivities and amenities.

In Heblich et al. [2025], we use these properties to extract industry-specific productivities across British cities and identify Jacobs externalities.



The economic literature has interpreted Jacobs' thesis as describing positive technological externalities *across* industries:

> *"[V]ariety and diversity of geographically proximate industries rather than geographical specialization promote innovation and growth."*
> *Glaeser et al. [1992]*

How should we proceed?

How should we proceed?

One needs to specify a parametric formulation for these externalities. The traditional interpretation of industry-specific dynamics has a shift-share flavor:

$$\mathcal{T}_{i\ell t} = \mathcal{A}_{i\ell} \cdot \mathcal{P}_{it}$$

where productivity is jointly determined by:

- shares: the distribution of location advantages, $\{\mathcal{A}_{j\ell}\}_{j \in I}$,
- shifts: industry-wide demand/technology shifters, $\{\mathcal{P}_{jt}\}_{j \in I}$.

We will assume instead (period 2 is 2020; period 1 is 1881):

$$\mathcal{T}_{i\ell 2} = \mathcal{A}_{i\ell} \cdot \mathcal{P}_{i2} \cdot L_{i\ell 2}^{\mu} \cdot L_{\ell 1}^{\rho} \cdot \left[\sum_{j \in I} \left(\frac{L_{j\ell 1}}{L_{\ell 1}} \right)^2 \right]^{-\lambda}$$

The empirical equivalent of the previous equation is,

$$\ln\left(\mathcal{T}_{i\ell 2}\right) = \mu l_{i\ell 2} + \rho l_{\ell 1} - \lambda h_{\ell 1} + \eta_i + \gamma \mathbf{X}_\ell + \varepsilon_{i\ell 2}, \tag{1}$$

where:

- population ($l_{\ell 1} = \ln(L_{\ell 1})$) and concentration ($h_{\ell 1} = \ln\left[\sum_{j \in I}\left(\frac{L_{j\ell 1}}{L_{\ell 1}}\right)^2\right]$) are observed in the data;
- $\eta_i = \ln(\mathcal{P}_{i2})$ captures contemporary industry shifters;
- our approach to isolating the fundamentals of city ℓ, $\ln \mathcal{A}_{i\ell}$, conditions the specification on controls capturing first-/second-nature geography, etc.

The empirical equivalent of the previous equation is,

$$\ln\left(\mathcal{T}_{i\ell 2}\right) = \mu l_{i\ell 2} + \rho l_{\ell 1} - \lambda h_{\ell 1} + \eta_i + \gamma \mathbf{X}_\ell + \varepsilon_{i\ell 2}, \tag{1}$$

where:

- population ($l_{\ell 1} = \ln(L_{\ell 1})$) and concentration ($h_{\ell 1} = \ln\left[\sum_{j \in I}\left(\frac{L_{j\ell 1}}{L_{\ell 1}}\right)^2\right]$) are observed in the data;
- $\eta_i = \ln(\mathcal{P}_{i2})$ captures contemporary industry shifters;
- our approach to isolating the fundamentals of city ℓ, $\ln \mathcal{A}_{i\ell}$, conditions the specification on controls capturing first-/second-nature geography, etc.

How would you identify λ?

To isolate exogenous variation in industrial concentration, we rely on identification within shift-share designs:

- endogenous shares: the distribution of location advantages, $\{\mathcal{A}_{j\ell}\}_{j\in J}$,
- exogenous shifts: industry-wide demand shifters, $\{\mathcal{P}_{j1}\}_{j\in J}$.

We proxy location advantages in city ℓ with employment shares in 1817, $s_{j\ell 0} = L_{j\ell 0}/L_{c0}$, and combine them with "external demand" across sectors:

$$\chi_\ell = \frac{\sum_j (s_{j\ell 0} \cdot d_j)^2}{\sum_j (s_{j\ell 0})^2} = \sum_j \sigma_{j\ell 0} \cdot d_j^2$$

where $\left\{\sigma_{j\ell 0} = (s_{j\ell 0})^2 / \sum_j (s_{j\ell 0})^2\right\}_{j\in J}$ are shares that sum up to 1.

Identification requires the shifts to be numerous and quasi-random:

- leave-out measures of aggregate sectoral employment growth,
- changes in *real input costs* between 1784–1896,
- changes in U.S. imports between 1824–1862 as a *demand shifter*.

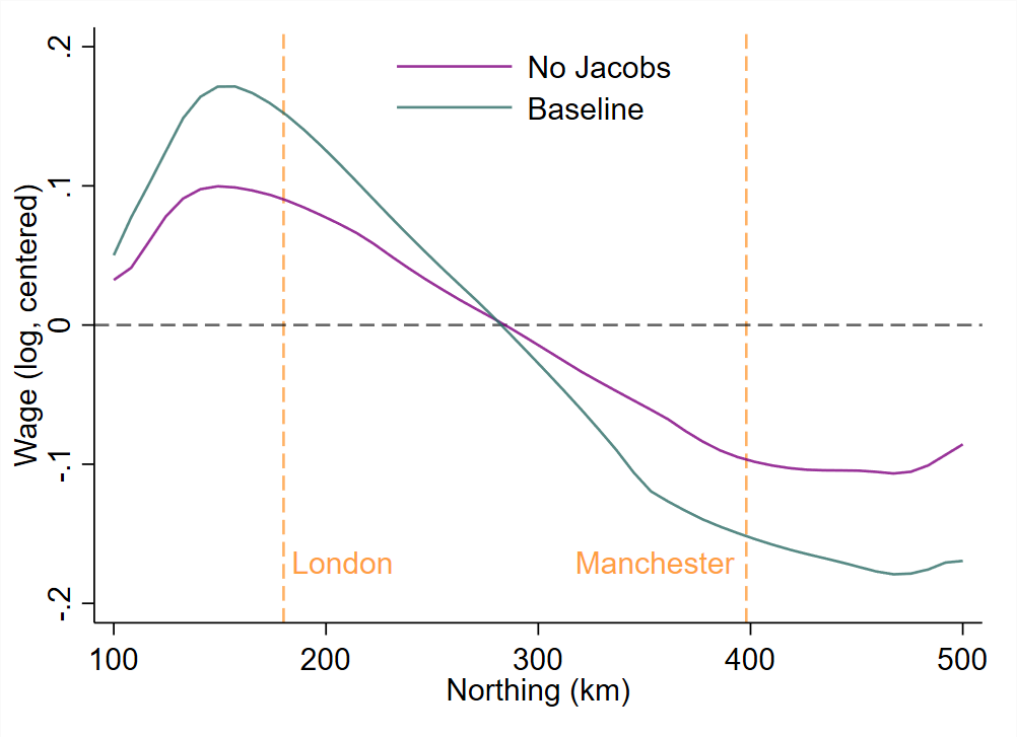
One can also control for a *linear* shift-share variable, $g_\ell = \sum_j s_{j\ell 0} \cdot d_j$.

$$\ln\left(\mathcal{T}_{i\ell 2}\right) = \mu l_{i\ell 2} + \rho l_{\ell 1} - \lambda h_{\ell 1} + \eta_i + \gamma \mathbf{X}_{\ell} + \varepsilon_{i\ell 2}$$

Recovered productivity ($\ln \mathcal{T}_{i\ell 2}$)	(1)	(2)
Herfindahl index ($h_{\ell 1} = \ln \sum_j \left(L_{j\ell 1}/L_{\ell 1}\right)^2$)	-1.227	-0.794
	(0.484)	(0.392)
	{18.40}	{18.56}
Employment ($l_{\ell 1} = \ln L_{\ell 1}$)	0.369	-0.030
	(0.437)	(0.341)
	{22.63}	{22.54}
Observations	38,280	38,280
Local industrial employment	No	Yes

These estimates imply a gap of about 14% in long-run productivity between quite diverse versus specialized cities (separated by one SD).

While the initial disparities between Northern and Southern England are around 30-40% in income...



Notes: This figure displays the best non-linear fit for wage as a function of "Northing." For the sake of exposition, we display the Northing levels of London and Manchester as dashed (orange) lines. The baseline is represented by the green line; the counterfactual is the purple line.

...neutralizing Jacobs externalities ($\lambda = 0$) would halve this gap.

Take-away

In spatial economics, elasticities disciplining location choice (**1st example**), production, housing construction, spillovers (**2nd example**), etc., are key to redistributing workers and firms across space.

They are, however, often calibrated rather than **credibly estimated**.

The previous examples have shown that model structure creates:

- challenges for the empirical estimation of key parameters,
- but also opportunities to discipline identification.

Can recent econometric advances help us credibly estimate spatial relationships such as elasticities and externalities?

Spatial identification

Consider a hedonic approach in a **large economy** [Rosen, 1974]:

$$u\left(w-p,b\right)=\mathcal{U}$$

Following a shock in amenity, we can differentiate with respect to b:

$$u_w\cdot\left(\frac{\partial w}{\partial b}-\frac{\partial p}{\partial b}\right)+u_b=\mathbf{0}$$

and thus,

$$\frac{\partial w}{\partial b}-\frac{\partial p}{\partial b}=-u_b/u_w$$

If we assume that wages are inelastic ($\partial w/\partial b=0$), then the change in house prices captures the standard willingness-to-pay,

$$\frac{\partial p}{\partial b}=u_b/u_w$$

If, instead, local prices are inelastic ($\partial p/\partial b=0$), then we have a standard compensating differential equation,

$$\frac{\partial w}{\partial b}=-u_b/u_w$$

In the first case, how can we estimate the willingness-to-pay (u_b/u_w) where b captures, for instance, urban heat during Summer?

In the first case, how can we estimate the willingness-to-pay (u_b/u_w) where b captures, for instance, urban heat during Summer?

a. a two-way fixed effect regression?

$$p_{\ell t} = \beta h_{\ell t} + \eta_\ell + \nu_t + \gamma \mathbf{X}_{\ell t} + \varepsilon_{\ell t}$$

What is the identification hypothesis? What are the threats?

In the first case, how can we estimate the willingness-to-pay (u_b/u_w) where b captures, for instance, urban heat during Summer?

a. a two-way fixed effect regression?

$$p_{\ell t} = \beta h_{\ell t} + \eta_\ell + \nu_t + \gamma \mathbf{X}_{\ell t} + \varepsilon_{\ell t}$$

What is the identification hypothesis? What are the threats?

b. a spatial regression discontinuity with the following identification hypothesis?

$$lim_{s \to \bar{s}} E\left[p_{\ell t} | s\right] = E\left[p_{\ell t} | \bar{s}\right], \quad t \in \{0, 1\}$$

In the first case, how can we estimate the willingness-to-pay (u_b/u_w) where b captures, for instance, urban heat during Summer?

a. a two-way fixed effect regression?

$$p_{\ell t} = \beta h_{\ell t} + \eta_\ell + \nu_t + \gamma \mathbf{X}_{\ell t} + \varepsilon_{\ell t}$$

What is the identification hypothesis? What are the threats?

b. a spatial regression discontinuity with the following identification hypothesis?

$$lim_{s \to \bar{s}} E\left[p_{\ell t} | s\right] = E\left[p_{\ell t} | \bar{s}\right], \quad t \in \{0, 1\}$$

c. a spatial diffusion process leading to heat dispersion/retention?

2.a. Difference-in-differences, SUTVA and general equilibrium

Spatial differences-in-differences are:

- differences-in-differences (“parallel trends” assumptions, homogeneous treatment effects),
- where one difference is across (sometimes nearby) locations.

Threats: movements of factors and goods, other treatment spillovers (SUTVA), heterogeneous effects and staggered adoption.

References (general): Baker et al. [2025].

References (spatial): Banzhaf [2021] (SUTVA); Butts [2021] (spillovers, among others), Arkhangelsky et al. [2021] (synthetic).

2.b. Spatial discontinuity

A spatial regression discontinuity is:

- a regression discontinuity (“continuity” assumption),
- where the forcing variable is the “spatial” distance to a border or a two-dimensional forcing variable (with a two-dimensional cutoff).

Threats: movements of factors and goods, other treatment spillovers (SUTVA).

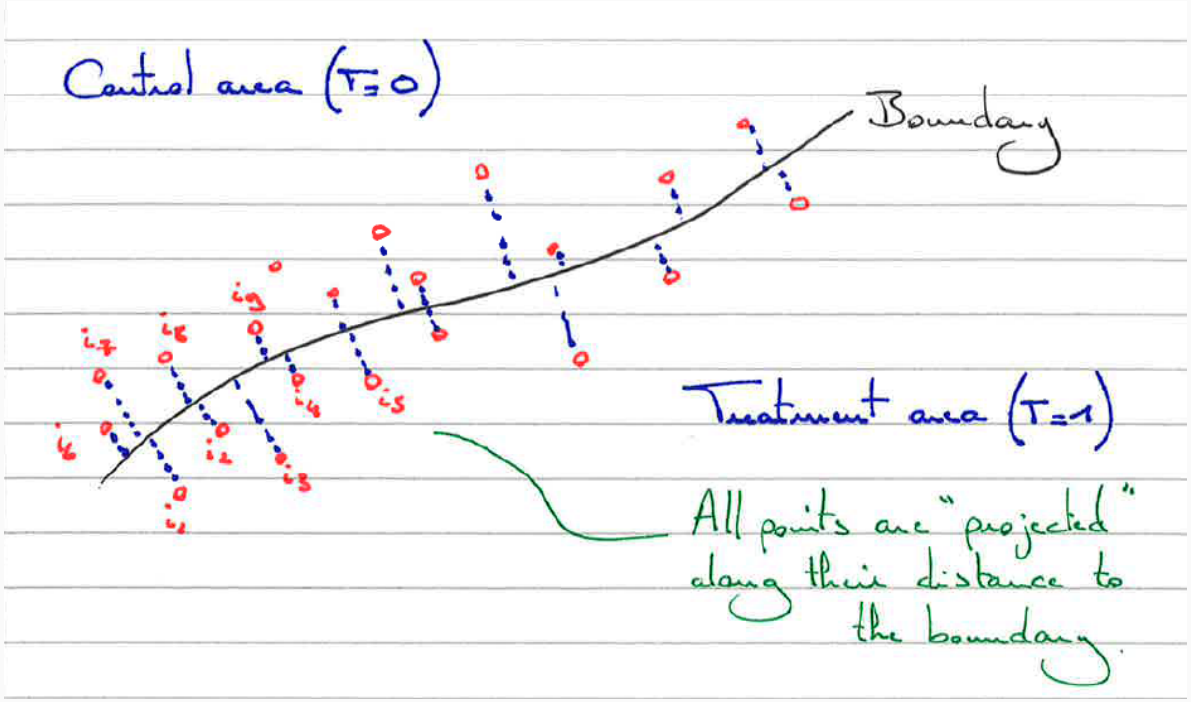
References: Imbens and Zajonc [2011]; Skovron and Titiunik [2015]; Keele and Titiunik [2015]; Calonico et al. [2014].

$\mathbf{A_1}$: Continuity in one-dimensional score, e.g., distance:

$$\lim_{d \to 0} E(y_{\ell,0}|d) = E(y_{\ell,0}|0), \quad \lim_{d \to 0} E(y_{\ell,1}|d) = E(y_{\ell,1}|0)$$

Under $\mathbf{A_1}$,

$$\lim_{d \to 0^+} E(y_\ell) - \lim_{d \to 0^-} E(y_\ell) = E(y_{\ell,1} - y_{\ell,0}|0)$$



Estimation under $\mathbf{A_1}$:

$$y_\ell = \beta t_\ell + P_0(d_\ell) + P_1(d_\ell) + \varepsilon_\ell$$

where:

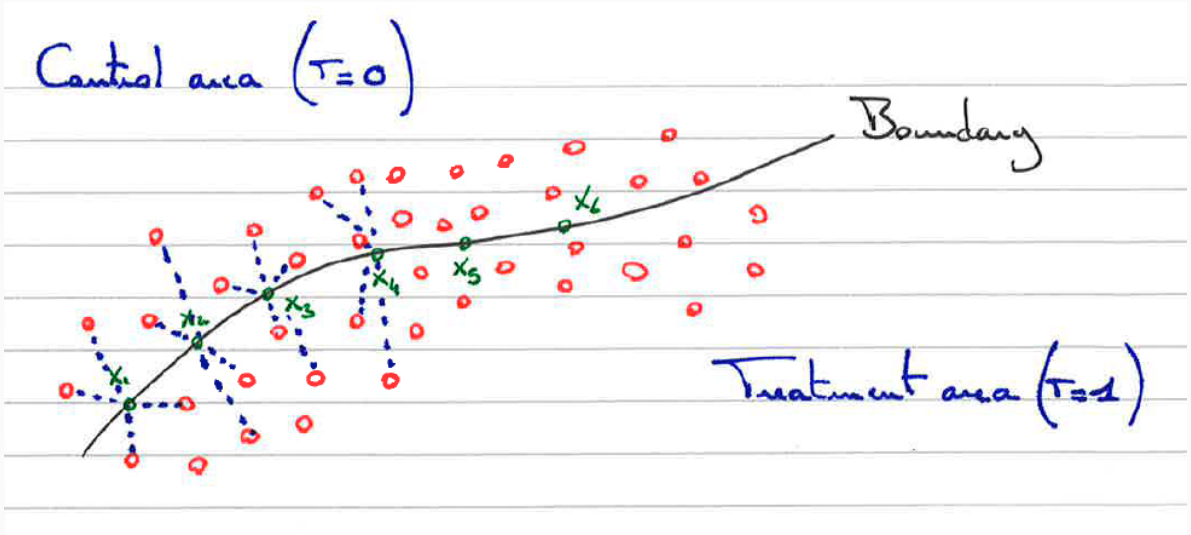
- y_ℓ is the outcome of interest measured for ℓ,
- d_ℓ is the distance to the cutoff for ℓ,
- t_ℓ is the treatment, e.g., 1 on one side versus 0 on the other side,
- P_τ are polynomials in d_ℓ, conditional on $t_\ell = \tau$.

A_2: Continuity in two-dimensional score, e.g., coordinates:

$$\lim_{(x_1,x_2)\to(\bar{x}_1,\bar{x}_2)} E(y_{\ell,t}|(x_1,x_2)) = E(y_{\ell,t}|(\bar{x}_1,\bar{x}_2)), \quad t = 0,1$$

Under **A_2**:

$$\lim_{(x_1,x_2)\in A_t\to(\bar{x}_1,\bar{x}_2)} E(y_\ell) - \lim_{(x_1,x_2)\in A_c\to(\bar{x}_1,\bar{x}_2)} E(y_\ell) = E(y_{\ell,1} - y_{\ell,0}|(\bar{x}_1,\bar{x}_2))$$



Estimation under $\mathbf{A_2}$:

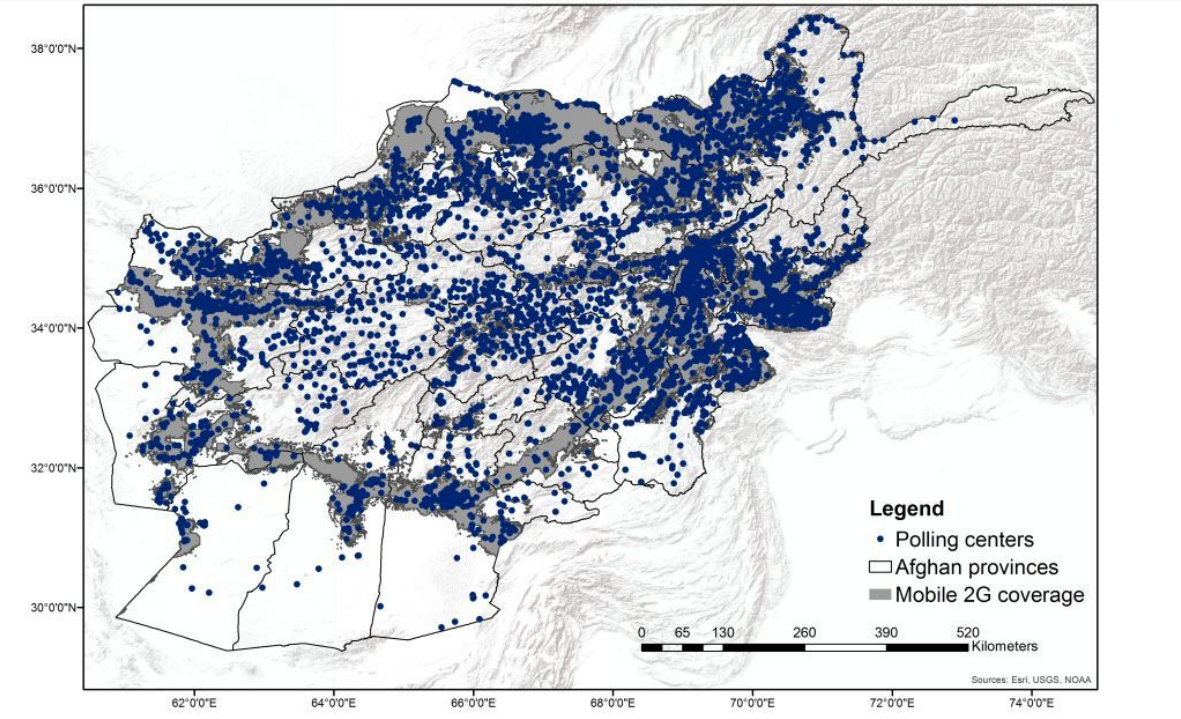
$$y_\ell = \beta_n t_\ell + P_{0,n}(\mathbf{x}_\ell) + P_{1,n}(\mathbf{x}_\ell) + \mu_n + \varepsilon_\ell$$

where:

- y_ℓ is the outcome of interest measured for ℓ,
- $(\mathbf{x}_\ell)$ are spatial coordinates for ℓ,
- t_ℓ is the treatment, e.g., 1 on one side versus 0 on the other side,
- P's are polynomials in $\mathbf{x}_\ell$.

Examples:

- historical borders [Dell, 2010; Michalopoulos and Papaioannou, 2014];
- electoral districts [Keele et al., 2015];
- school district boundaries [Black, 1999];
- terrain discontinuities [Gonzalez, 2016].



2.c. Diffusion processes

A novel approach to identification is the use of spatial diffusion processes or non-random exposure approaches [Borusyak and Hull, 2023].

Consider the empirical model,

$$y_\ell = \beta t_\ell + \varepsilon_\ell$$

where t_ℓ is the exposure of location ℓ to a certain treatment.

One can use a modeling of connections through space to explain treatment as a diffusion process, e.g.,

$$z_\ell = f_\ell(\omega)$$

Examples:

- a traditional shift-share trade shock,

$$z_\ell = \sum_{j \in \mathcal{J}} a_{j\ell} p_j$$

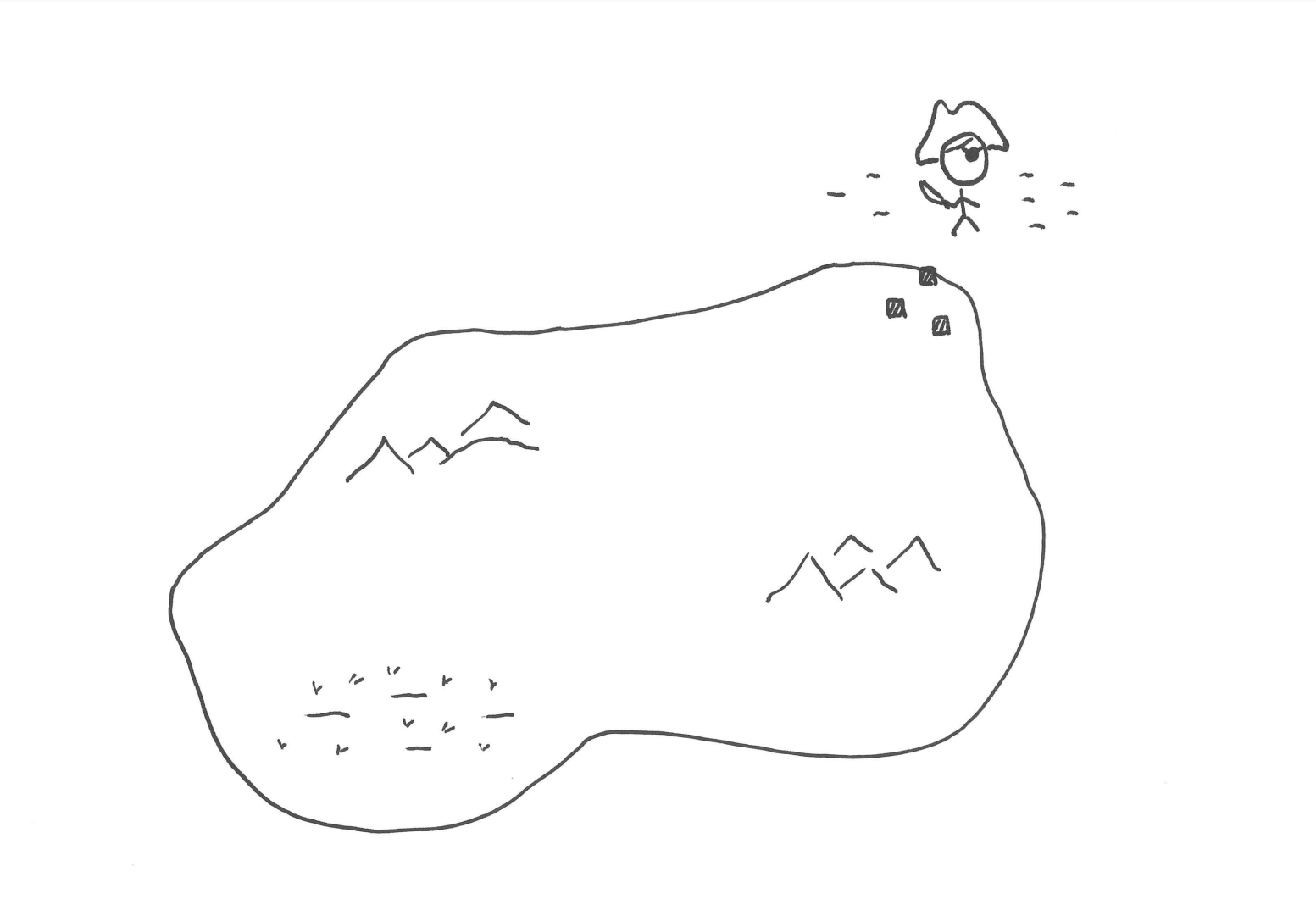
- the movement of certain individuals across space (e.g., rural migrants),

$$z_\ell = \frac{v_\ell^\chi}{\sum_{\ell' \in \mathcal{L}} (v_{\ell'})^\chi}$$

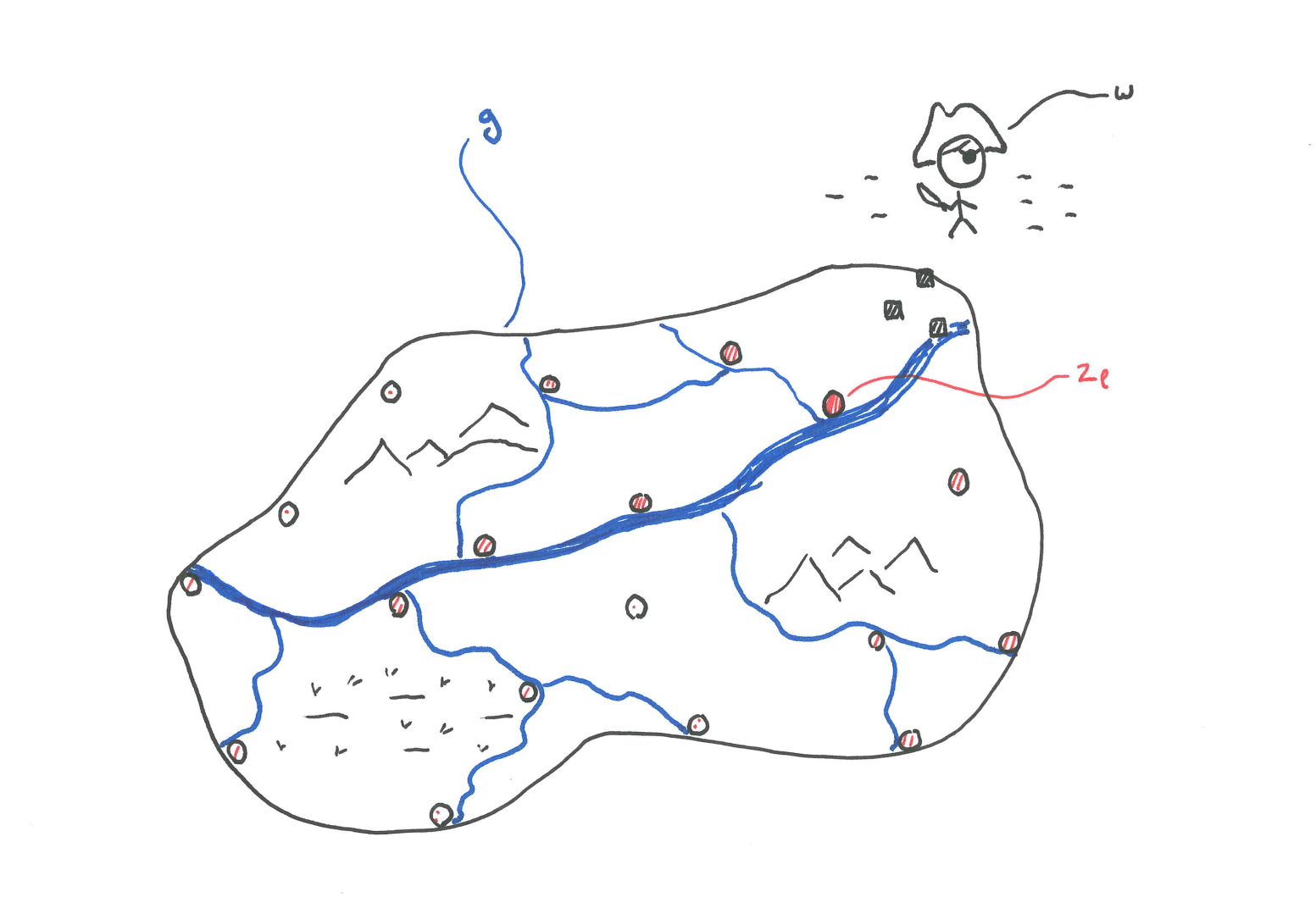
- the exposure to a shock g through the railway network, or to pollution emissions g through a wind dispersion process,

$$z_\ell = f_\ell(g, w)$$

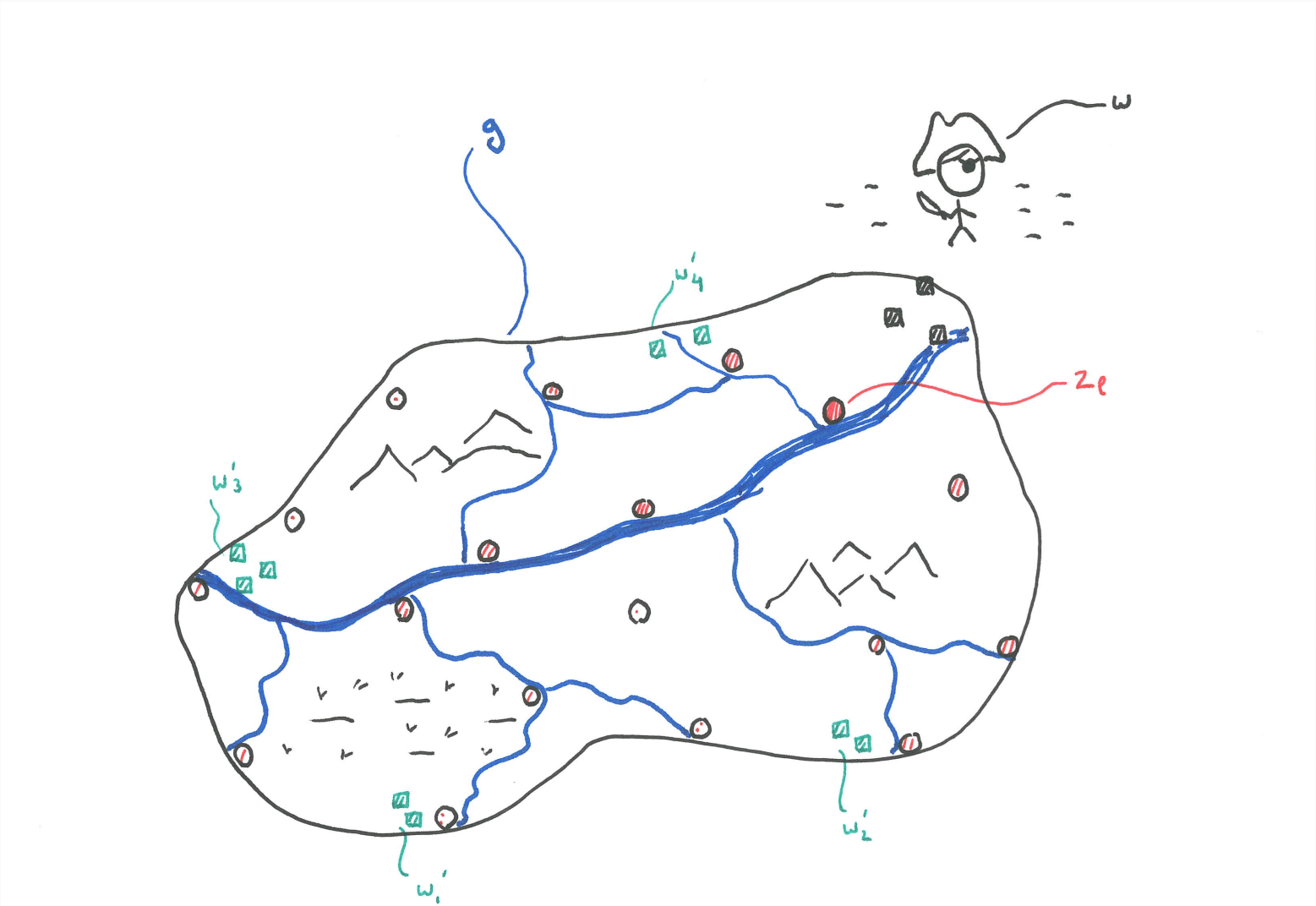
Many papers use such intuition without formalizing it [radio/TV antennas, see Olken, 2009; Yanagizawa-Drott, 2014].



Notes: An island, and a pirate with a Bristolian accent.



Notes: An island, and a pirate with a Bristolian accent. Blue line: roads, circles: locations of interest, shades of red: treatment.



Notes: An island, a pirate (w) with a Bristolian accent, alternative pirates (w'). Blue line: roads, circles: locations of interest, shades of red: treatment.

Identification with non-random spatial diffusion processes [Borusyak and Hull, 2023]:

$$y_\ell = \beta t_\ell + \varepsilon_\ell$$

where t_ℓ, the exposure of location ℓ to the treatment, is instrumented by:

$$z_\ell = f_\ell(g, w)$$

and w are exogenous shocks. The instrument needs to be re-centered through the construction of an expected instrument:

$$\mu_\ell = E\left[f_\ell(g, w)|w\right]$$

using alternative shocks w' drawn from a reasonable distribution.

Notes: in the shift-share case, re-centering is not needed due to linearity and the fact that the shares sum up to 1 for each location.

2.d. Inference

What about inference (with spatial econometrics)?

- heteroskedastic, continuous dependence [Conley, 1999; Müller and Watson, 2024];
- clustering [Barrios et al., 2012; Abadie et al., 2017];
- shift-share designs [Borusyak et al., 2025];
- regression-discontinuity designs [Calonico et al., 2014];
- or non-random exposure approaches [Borusyak and Hull, 2023].

3. Questions?

Thank you very much!

yanos.zylberberg@bristol.ac.uk

Appendix

References

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J.: 2017, When should you adjust standard errors for clustering?, *Technical report*, National Bureau of Economic Research.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W. and Wager, S.: 2021, Synthetic difference-in-differences, *American Economic Review* **111**(12), 4088–4118.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A. and Sant'Anna, P. H.: 2025, Difference-in-differences designs: A practitioner's guide, *arXiv preprint arXiv:2503.13323* .

Banzhaf, H. S.: 2021, Difference-in-differences hedonics, *Journal of Political Economy* **129**(8), 2385–2414.

Barrios, T., Diamond, R., Imbens, G. W. and Kolesár, M.: 2012, Clustering, spatial correlations, and randomization inference, *Journal of the American Statistical Association* **107**(498), 578–591.

Black, S. E.: 1999, Do better schools matter? parental valuation of elementary education, *The Quarterly Journal of Economics* **114**(2), 577–599.

Borusyak, K. and Hull, P.: 2023, Nonrandom exposure to exogenous shocks, *Econometrica* **91**(6), 2155–2185.

Borusyak, K., Hull, P. and Jaravel, X.: 2025, A practical guide to shift-share instruments, *Journal of Economic Perspectives* **39**(1), 181–204.

Butts, K.: 2021, Difference-in-differences estimation with spatial spillovers, *arXiv preprint arXiv:2105.03737* .

Calonico, S., Cattaneo, M. D. and Titiunik, R.: 2014, Robust nonparametric confidence intervals for regression-discontinuity designs, *Econometrica* **82**(6), 2295–2326.

Conley, T. G.: 1999, Gmm estimation with cross sectional dependence, *Journal of econometrics* **92**(1), 1–45.

Dell, M.: 2010, The persistent effects of peru's mining mita, *Econometrica* **78**(6), 1863–1903.

Gonzalez, R. M.: 2016, Social monitoring and electoral fraud: Evidence from a spatial regression discontinuity design in afghanistan, *Technical report*.

Heblich, S., Nagy, D. K., Trew, A. and Zylberberg, Y.: 2025, The death and life of great british cities, *Technical report*, National Bureau of Economic Research.

Imbens, G. and Zajonc, T.: 2011, Regression discontinuity design with multiple forcing variables, *Report, Harvard University.[972]* .

Imbert, C., Monras, J., Seror, M., Zylberberg, Y. et al.: 2023, Floating population: migration with (out) family and the spatial distribution of economic activity.

Keele, L. J. and Titiunik, R.: 2015, Geographic boundaries as regression discontinuities, *Political Analysis* **23**(1), 127–155.

Keele, L., Titiunik, R. and Zubizarreta, J. R.: 2015, Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout, *Journal of the Royal Statistical Society: Series A* **178**(1), 223–239.

Michalopoulos, S. and Papaioannou, E.: 2014, National institutions and subnational development in africa, *The Quarterly Journal of Economics* **129**(1), 151–213.

Müller, U. K. and Watson, M. W.: 2024, Spatial unit roots and spurious regression, *Econometrica* **92**(5), 1661–1695.

Olken, B. A.: 2009, Do television and radio destroy social capital? evidence from indonesian villages, *American Economic Journal: Applied Economics* **1**(4), 1–33.

Rosen, S.: 1974, Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of political economy* **82**(1), 34–55.

Skovron, C. and Titiunik, R.: 2015, A practical guide to regression discontinuity designs in political science.

Yanagizawa-Drott, D.: 2014, Propaganda and conflict: Evidence from the rwandan genocide, *The Quarterly Journal of Economics* **129**(4), 1947–1994.