Spatial Economics for Granular Settings

颗粒状设定下的空间经济学

Jonathan I. Dingel & Felix Tintelnot

(More revisions requested at *Econometrica*, September, 2023)

程芸倩　陈泽宇

中国经济转型讨论班 (CETW)

2024–05–14

目录

空间经济学的前沿趋势

- 现象 1: 经济活动呈现高度集聚特征
- 现象 2: 细颗粒度数据越来越可得，可以探究小空间尺度上发生的经济现象 (如: 卫星数据和手机数据)
- 现象 3: 地区对层面的“流”数据，用于探究地区间的经济联系 (如: 通勤数据)
- 何为细颗粒度 (granular)?
  - 柏林: 300 万通勤人口、2.54 亿个街区对 (Ahlfeldt et al., 2015)
  - 底特律城市地区: 130 万通勤人口、130 万个区块对 (Owens et al., 2020)
  - 洛杉矶都会区: 670 万通勤人口、600 万个区块对 (Severen, 2021)

当传统模型遇到细颗粒数据会有什么问题?

- 细颗粒度数据的一个特征: 数据中大面积的 0 或 0-5
- 试想在通勤的情境下
  - 规律性因素和偶然性因素
  - 大样本数据下: 规律性因素 dominate
  - 细颗粒度数据下: 偶然性因素 dominate

当传统模型遇到细颗粒数据会有什么问题?

- 细颗粒度数据的一个特征: 数据中大面积的 0 或 0-5
- 细颗粒度数据下: 偶然性因素 dominate
- 偶然因素对传统 QSM 模型的挑战
  - 传统模型理论推导中的假设需要依靠大数定律 vs 颗粒状设定下很难认为观测数据是大数定律作用过后得出的结果
  - 传统模型反事实结果估计了点估计值 vs 颗粒状设定下反事实结果是不确定的, 依赖于个体决策的“偶然性因素”
- 细颗粒度设定下, 使用传统 QSM 模型对结果的挑战
  - 传统 QSM 模型: 试图 match all data
  - 模型会过度拟合了数据中的偶然因素, 产生“过拟合问题”
  - 需要提出适应细颗粒度设定的新模型

结构估计运作方式简述

一个抽象的空间经济学模型：

结构估计运作方式简述

Environment:

- # of regions: N
- ...

Endowments:

- # of total population: L
- # of lands by regions: $\{T_k\}_{k=1}^{N}$
- Productivity by regions: $\{A_n\}_{n=1}^{N}$
- ...

Bilateral resistance:

- Migration/commuting costs: $\{\delta_{kn}\}_{k=1,n=1}^{N,N}$
- Trade costs: $\{0\}_{k=1,n=1}^{N,N}$
- ...

...

Distribution of population: $\{l_{kn}\}_{k=1,n=1}^{N,N}$

Wage by regions: $\{w_n\}_{n=1}^{N}$

Land price by regions: $\{r_k\}_{k=1}^{N}$

Welfare

...

结构估计运作方式简述

Exogenous variables —*input*→ ***Optimization*** / ***Market clearing*** —*output*→ ***Endogenous variables***

Environment:

- # of regions: N
- ...

Endowments:

- # of total population: L
- # of lands by regions: $\{T_k\}_{k=1}^{N}$
- Productivity by regions: $\{A'_n\}_{n=1}^{N}$
- ...

Bilateral resistance:

- Migration/commuting costs: $\{\delta_{kn}\}_{k=1,n=1}^{N,N}$
- Trade costs: $\{0\}_{k=1,n=1}^{N,N}$
- ...

...

Distribution of population: $\{l'_{kn}\}_{k=1,n=1}^{N,N}$

Wage by regions: $\{w'_n\}_{n=1}^{N}$

Land price by regions: $\{r'_k\}_{k=1}^{N}$

Welfare

...

结构估计运作方式简述

Environment:
- # of regions: N
- ...

Endowments:
- # of total population: L
- # of lands by regions: $\{T_k\}_{k=1}^{N}$
- Productivity by regions: $\{A'_n\}_{n=1}^{N}$
- ...

Bilateral resistance:
- Migration/commuting costs: $\{\delta_{kn}\}_{k=1,n=1}^{N,N}$
- Trade costs: $\{0\}_{k=1,n=1}^{N,N}$
- ...

...

Distribution of population: $\{l'_{kn}\}_{k=1,n=1}^{N,N}$
Wage by regions: $\{w'_n\}_{n=1}^{N}$
Land price by regions: $\{r'_k\}_{k=1}^{N}$
Welfare
...

Decomposition of structural estimation:

1. 校准参数;
2. 观测(并校准)基准均衡(冲击前均衡);
3. 计算政策使某一/某些外生变量发生了多大变化;
4. 保持其他外生变量不变，将反事实(冲击后)的外生变量代入结构模型，求解反事实均衡(冲击后均衡);
5. 比较基准均衡与反事实均衡下感兴趣的内生变量的变化，求得政策效应。

结构估计的运作：校准基准均衡的直觉阐释

- 能够用直接完全使用观测数据代表基准均衡吗？
  - 不可以！结构模型会推导出各变量在均衡下的一系列等式关系，但这些变量的观测数据不可能恰好满足这些关系。
- 比如说，均衡时如下式子成立：

$$A = B + C$$

我们有变量 A、B 和 C 的观测数据，但显然不可能恰好满足该均衡关系。

  - 处理方法之一：放弃变量 A 的观测数据，保留变量 B 和 C 的观测数据，用 B 和 C 计算出 A，以此作为 A 的基准均衡取值。
  - 为什么是放弃 A 而不是 B 呢？放弃 C 呢？
  - 处理思路：放弃那个与结构模型相比，含义最不一样的观测变量。

目录

基本设定

- 静态模型 (只有一期)，完全竞争。
- 封闭经济体，共有 L 单位劳动力 (居民) 和 N 个地区 (下标 k、n)。
- 每个地区有固定数量的土地 T_k，只用于居民住房消费；土地收入归当地地主所有，并转化为消费。
- 每个地区生产一种特有的商品，无贸易成本。
- 劳动力选择居住地与工作地；居住地与工作地如果不同，需要承担通勤成本 δ_{kn}，用冰川成本表示：

$$\delta_{kn} \equiv \underbrace{\bar{\delta}_{kn}}_{\text{可观测成本: } f(\text{通勤时间})} \times \underbrace{\lambda_{kn}}_{\text{不可观测成本}}$$

Optimazation

- 选择在地区 k 居住、在地区 n 工作的工人 i 的效用函数为：

$$u_{kn}^{i} = \epsilon \quad \underbrace{\ln C_{kn}}_{\text{地区对层面特征}} \quad + \quad \underbrace{\nu_{kn}^{i}}_{\text{个体异质性}}$$

- 消费指数 C_{kn} 取 Cobb-Douglas 函数的形式：

$$C_{kn} = \frac{1}{\lambda_{kn}} \left(\frac{c_{kn}}{1-\alpha} \right)^{1-\alpha} \left(\frac{T_{kn}}{\alpha} \right)^{\alpha}$$
$$\text{s.t. } P c_{kn} + r_k T_{kn} = w_n / \bar{\delta}_{kn}$$

其中，最终品 c_{kn} 是各地生产商品的 CES 加总 (商品间替代弹性为 $\sigma > 1$)，T_{kn} 是对住房 (土地) 的消费，最终品价格为 $P = \left[\sum_n (p_n)^{1-\sigma} \right]^{1/(1-\sigma)}$，土地价格为 r_k。

- 特异性偏好 ν_{kn}^{i} 是一个随机变量，i.i.d 地服从于一个 Gumbel 分布：

$$F(\nu_{kn}^{i}) = e^{-e^{-\nu_{kn}^{i}}}$$

Optimazation

- 通过最大化上述效用函数，可得居民的间接效用函数为：

$$U_{kn}^{i} = \epsilon \ln\left(\frac{w_n}{r_k^{\alpha} P^{1-\alpha} \delta_{kn}}\right) + \nu_{kn}^{i} \tag{1}$$

- 各地的生产函数：

$$q_n = A_n L_n$$

其中 A_n 为地区生产率，L_n 为地区劳动力。

$$\Rightarrow \text{单位产出成本 (商品价格): } p_n = w_n / A_n$$

$$\Rightarrow \text{价格指数: } P = \left[\sum_n \left(\frac{w_n}{A_n}\right)^{1-\sigma}\right]^{1/(1-\sigma)}$$

Optimazation

- 工人选择居住地与工作地来实现效用最大化，由于存在随机项 $\{\nu_{kn}^i\}_{k=1,n=1}^{N,N}$，选择在地区 k 居住、在地区 n 工作是一个概率：

$$m_{kn} = \Pr\left\{\epsilon \ln\left(\frac{w_n}{r_k^\alpha P^{1-\alpha}\delta_{kn}}\right) + \nu_{kn}^i \geq \max_{k',n'}\left\{\epsilon \ln\left(\frac{w_{n'}}{r_{k'}^\alpha P^{1-\alpha}\delta_{k'n'}}\right) + \nu_{k'n'}^i\right\}\right\}$$

- 代入 Gumbel 分布的分布函数，当工人数量足够大的时候，在地区 k 居住、在地区 n 工作的工人比例将收敛于概率 m_{kn}：

$$\frac{l_{kn}}{L} = \frac{w_n^\epsilon \left(r_k^\alpha \delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^\epsilon \left(r_{k'}^\alpha \delta_{k'n'}\right)^{-\epsilon}} \tag{2}$$

Market clearing

- 商品市场出清: 地区 n 产出 = 地区 n 需求

$$A_n \sum_k \frac{l_{kn}}{\bar{\delta}_{kn}} = \frac{(w_n/A_n)^\sigma}{P^{1-\sigma}} Y, \quad \forall n \tag{3}$$

其中 Y 是整个经济体的工资收入，即 $Y \equiv \sum_{k,n} y_{kn} \equiv \sum_{k,n} w_n l_{kn}/\bar{\delta}_{kn}$。

- 土地市场出清: 地区 k 土地供给 = 地区 k 土地需求

$$r_k T_k = \alpha \sum_n \frac{w_n}{\bar{\delta}_{kn}} l_{kn}, \quad \forall k \tag{4}$$

- 可以证明，当参数满足一定条件时 (即 $\left(\frac{1+\epsilon}{\sigma+\epsilon}\right)\left(\frac{\alpha\epsilon}{1+\alpha\epsilon}\right) \leq \frac{1}{2}$)，量化空间模型存在唯一的均衡，该均衡表现为一组工资水平、土地价格水平与人口分布的取值

$$\{w_n, r_k, l_{kn}\}_{k,n}$$

使得均衡条件式 (2) 至式 (4) 成立。

两种校准人口分布基准均衡的思路

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}}$$

- 参数与变量 (均具有现实观测):
  - α – 住房消费份额
  - ϵ – 效用函数中 consumption 相比起 idiosyncratic preference 的权重
  - l_{kn}/L – 居住在地区 k、工作在地区 n 的人口份额
  - w_n – 工作地 n 的工资
  - r_k – 居住地 k 的土地价格 (房价)
  - $\delta_{kn} \equiv \bar{\delta}_{kn} \times \lambda_{kn}$ – 通勤成本
- 貌似有两项的现实观测与模型 (均衡时) 的含义不太一致:
  - 通勤成本 $\delta_{kn} \equiv \bar{\delta}_{kn} \times \lambda_{kn}$: 利用观测数据只能建模出其中 $\bar{\delta}_{kn}$ 一项。
  - 人口份额 l_{kn}/L: 模型中的完整含义为“大样本理论作用下的人口份额”, 但观测数据未必反映大样本理论作用下的结果。

两种校准人口分布基准均衡的思路

协变量方法 (Covariates-Based Approach, CBA)

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}}$$

- 思路一: 已知 δ_{kn} 反推 l_{kn}/L
  - 利用可观测的协变量建模通勤成本 δ_{kn}(例如将其设定为两地通勤时间 τ_{kn} 的函数), 然后利用式 (2) 拟合人口份额。
  - 显然, 拟合出的人口份额和观测的人口份额肯定是不一致的, 但我们将拟合出的人口份额视为基准均衡的取值, 用于反事实估计 (plug the fitted model's values)
  - 其优点在于对数据的需求少 (比如无需知道现实中的人口份额), 但其潜在问题在于, 对通勤成本进行建模不得不遗漏掉不可观测的因素。

两种校准人口分布基准均衡的思路

精确帽代数 (Exact Hat Algebra, EHA)

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon} \left(r_k^{\alpha} \delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon} \left(r_{k'}^{\alpha} \delta_{k'n'}\right)^{-\epsilon}}$$

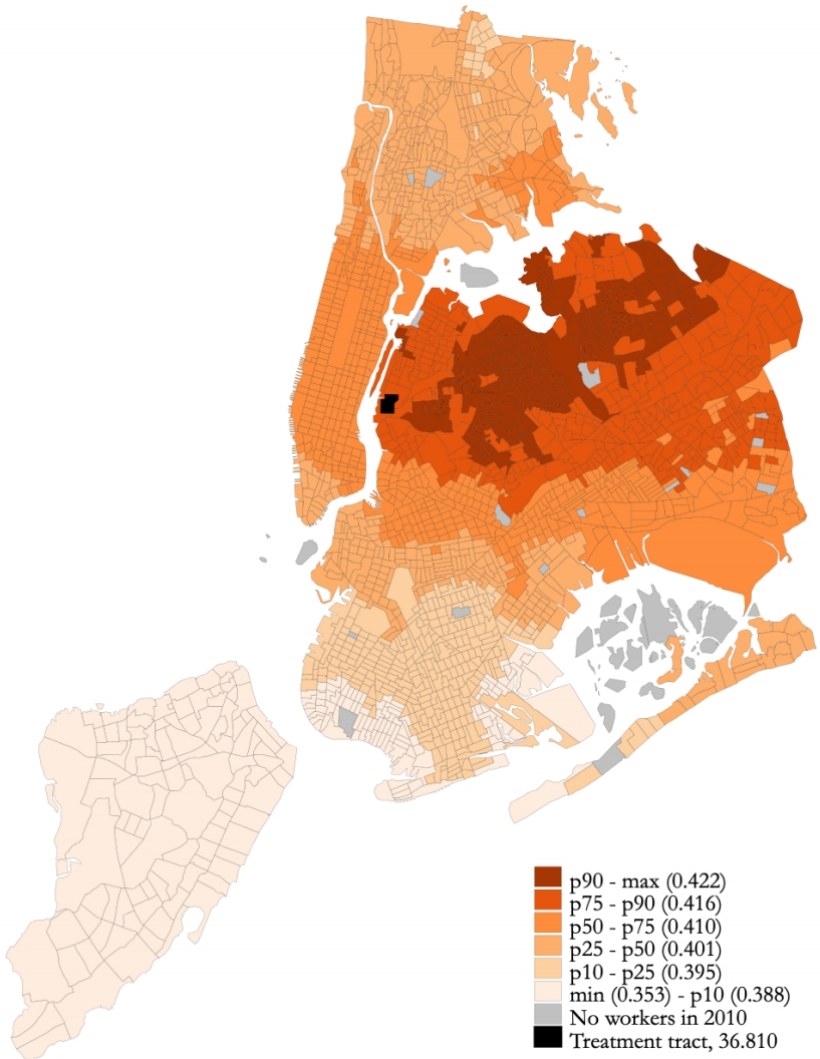
- 思路二：已知 l_{kn}/L 反推 δ_{kn}
  - 直接认为观测到的人口份额就是基准均衡，此时将观测数据代入式 (2)，拟合出通勤成本 δ_{kn}。
  - 在反事实估计中，直接代入观测到的人口份额 (plug the observed values)
  - 这一思路暗含了一个假设：现实的观测数据与连续模型基于大样本理论推导出的人口分布是接近的。如此一来，拟合出的 δ_{kn} 既包含了可观测成本又包含不可观测成本。然而，在颗粒状设定下，这个假设可能问题很大。

目录

纽约市区块间的通勤

- “区块”是指美国人口普查局 (U.S. Census Bureau) 定义的一个小区域，通常是一个相对连续的地理区域，用于收集和分析人口统计数据。
- 区块是一个空间尺度很小的单位，光是纽约一座城市就有大约 460 万个“区块对”，相比之下，纽约只有 250 万的通勤人口。
  - 85% 的区块对是没有人口的;
  - 在剩下的 15% 的区块对中，超过一半的区块对只有 1 个人。

纽约市区块间的通勤


p90 - max (0.422)
p75 - p90 (0.416)
p50 - p75 (0.410)
p25 - p50 (0.401)
p10 - p25 (0.395)
min (0.353) - p10 (0.388)
No workers in 2010
Treatment tract, 36.810

纽约市区块间的通勤

- 颗粒状设定下区块对的通勤人口数量是不稳定的，尤其是对于人口数量很少的区块对来说：
  - 那些在 2013 年有 1 个通勤人口的区块对，在 2014 年有 65% 变成了 0，只有 20% 仍保持 1 人。
  - 在 2013 年有 2 个通勤人口的区块对，也只有 15% 在 2014 年仍保持 2 人。
- 对于通勤人口数量很少的区块，人数主要的影响因素可能是一些随机因素。

(b) NYC

	2014					
2013	0	1	2	3	4	5+
0	0.91	0.07	0.01	0.00	0.00	0.00
1	0.65	0.20	0.08	0.04	0.02	0.02
2	0.39	0.25	0.15	0.09	0.05	0.07
3	0.24	0.22	0.17	0.12	0.08	0.16
4	0.15	0.17	0.17	0.14	0.11	0.27
5+	0.03	0.05	0.06	0.07	0.07	0.71

校准基准均衡

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}}$$

- 下面，我们基于 2010 年的现实数据，来校准模型的基准均衡。
- 可得的参数和变量：住房消费份额 $\alpha = 0.24$、观测到的人口份额 l_{kn}/L、可观测的通勤成本 $\bar{\delta}_{kn}$(、各地工资水平 w_n、各地的地价 r_k):
  - 通勤成本可以分为两项：

$$\delta_{kn} \equiv \underbrace{\bar{\delta}_{kn}}_{\text{可观测成本: } f(\text{通勤时间})} \times \underbrace{\lambda_{kn}}_{\text{不可观测成本}}$$

  - 利用 Google Map 计算出各区块来回的通勤时间 t_{kn} 和 t_{kn}，进而计算出 $\bar{\delta}_{kn} = H/\left(H - t_{kn} - t_{nk}\right)$，其中 H 代表工人一天的工作时间，取 $H = 9$。
- 下面我们先校准参数 ϵ!

校准基准均衡

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\Phi}$$

- 两边取对数：

$$\begin{aligned}\ln\left(\frac{l_{kn}}{L}\right) &= -\ln\Phi + \epsilon\ln w_n - \epsilon\alpha\ln r_k - \epsilon\ln\bar{\delta}_{kn} - \epsilon\ln\lambda_{kn} \\ &= \alpha + \eta_k + \psi_n - \epsilon\ln\bar{\delta}_{kn} + u_{kn}\end{aligned}$$

  - 识别假设：$\mathbb{E}\left[\ln\lambda_{kn}^{-\epsilon}|r_k, w_n, \ln\bar{\delta}_{kn}\right] = 0$，此时固定效应和 $\hat{\epsilon}$ 是一致估计量。
  - 根据两组固定效应可以得到工作地的工资 $\{w_n\}_n^N$ 和居住地的土地价格 $\{r_k\}_k^N$，因此即使没有观测数据也没事。
  - 由于 granular setting 下大量的 $l_{kn}/L = 0$，如果用“取对数 +OLS”会损失大量样本，所以估计的时候用的是其实是 PPML 而不是 OLS。

校准基准均衡

- 由于 OLS 损失了大量样本，以 MLE 的估计结果为准，ϵ 取 7.986。
- 估计中还得到了各居住地与工作地的固定效应，可以用固定效应推出各地的工资水平 w_n 和地价 r_k。

Table 1: Commuting elasticity estimates

	PPML/MLE	OLS
Commuting cost	-7.986	-2.307
	(0.307)	(0.0516)
Model fit (R^2 or pseudo-R^2)	0.662	0.561
Location pairs	4,628,878	690,673
Commuters	2,488,905	2,488,905

NOTES: All specifications include residence fixed effects and workplace fixed effects. The "PPML/MLE" column presents the results from maximum likelihood estimation of equation (8). The "OLS" column presents the results of estimating the log version of equation (2) by ordinary least squares, omitting observations in which $\ell_{kn} = 0$. The model-fit statistic is the pseudo-R^2 for MLE and R^2 for OLS. We report the PPML standard errors (clustered by k and by n), which are larger than the logit MLE standard errors associated with maximizing equation (8).

校准基准均衡

$$\frac{l_{kn}}{L} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}}$$

- 对于“协变量方法”而言，基准均衡下的人口份额是未知的，因此将校准出的 ϵ 代入式 (2)，将 l_{kn}/L 解出来。
- 对于“精确帽代数”而言，观测数据就是基准均衡下的人口份额。
- 完成上述以后，可以计算出基准均衡中各地区对的收入份额：

$$y_{kn} \equiv \frac{w_n}{\bar{\delta}_{kn}} l_{kn}$$

反事实分析

- 前面介绍结构估计的运作方式的时候说，直觉上，若要模拟某一冲击的经济影响，我们首先需要求解政策发生前的均衡 (基准均衡)，再求解政策发生后的均衡 (反事实均衡)，随后比较反事实均衡与基准均衡，从而得出政策的影响。
- 但处理上，其实有一些简化的方法。我们关注的其实只是冲击导致的变化量，而不是冲击前后的绝对量，因此我们可以将变量在反事实均衡 (x') 与基准均衡 (x) 的相对变化 $\hat{x} \equiv x'/x$ 作为估计的主要目标。
- 当估计目标转向相对变化的时候，很多变量或参数得以消去，因此能减少估计的难度 (Dekle et al., 2007, *AER*)。

反事实分析

- 将前面的均衡条件写成相对变化的形式:

$$\hat{w}_n = \hat{A}_n \left(\sum_k \hat{y}_{kn} \frac{y_{kn}}{\sum_{k'} y_{k'n}} \right)^{\frac{1}{1-\sigma}} \hat{P} \hat{Y}^{\frac{1}{\sigma-1}} \tag{5}$$

$$\hat{r}_k = \hat{T}_k^{-1} \sum_n \hat{y}_{kn} \frac{y_{kn}}{\sum_{n'} y_{kn'}} \tag{6}$$

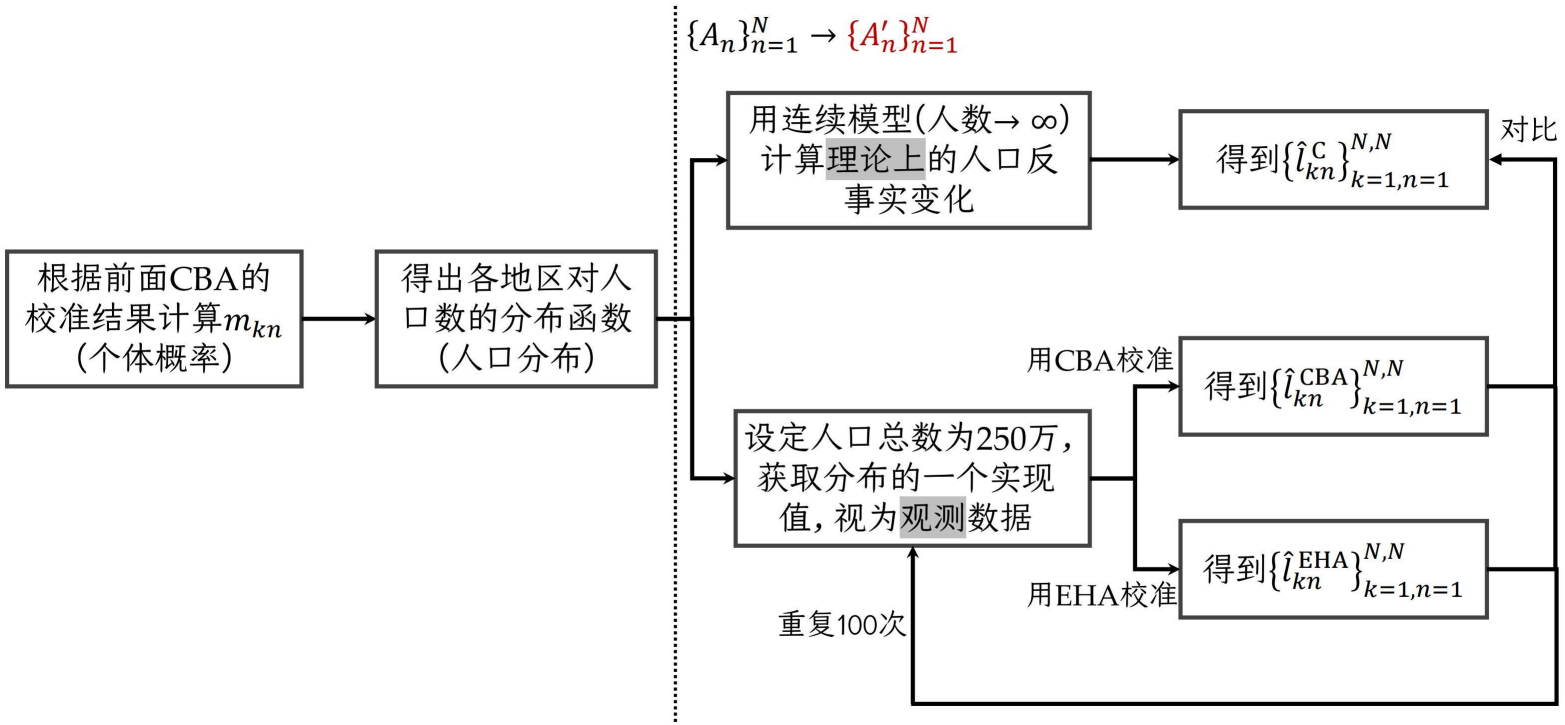
$$\hat{l}_{kn} = \frac{\hat{w}_n{}^{\epsilon} \left(\hat{r}_k{}^{\alpha} \hat{\bar{\delta}}_{kn} \hat{\lambda}_{kn} \right)^{-\epsilon}}{\sum_{k',n'} \hat{w}_{n'}{}^{\epsilon} \left(\hat{r}_{k'}{}^{\alpha} \hat{\bar{\delta}}_{k'n'} \hat{\lambda}_{k'n'} \right)^{-\epsilon} \frac{l_{k'n'}}{L}} \quad \text{if } l_{kn} > 0 \tag{7}$$

- 将人口和收入的基准份额、外生变量的反事实变化代入上面相对变化形式的均衡条件, 即可求得内生变量的反事实变化。

蒙特卡洛模拟: 两种思路的比较

$$\frac{l_{kn}}{L} = m_{kn} = \frac{w_n^{\epsilon}\left(r_k^{\alpha}\delta_{kn}\right)^{-\epsilon}}{\sum_{k',n'} w_{n'}^{\epsilon}\left(r_{k'}^{\alpha}\delta_{k'n'}\right)^{-\epsilon}}$$

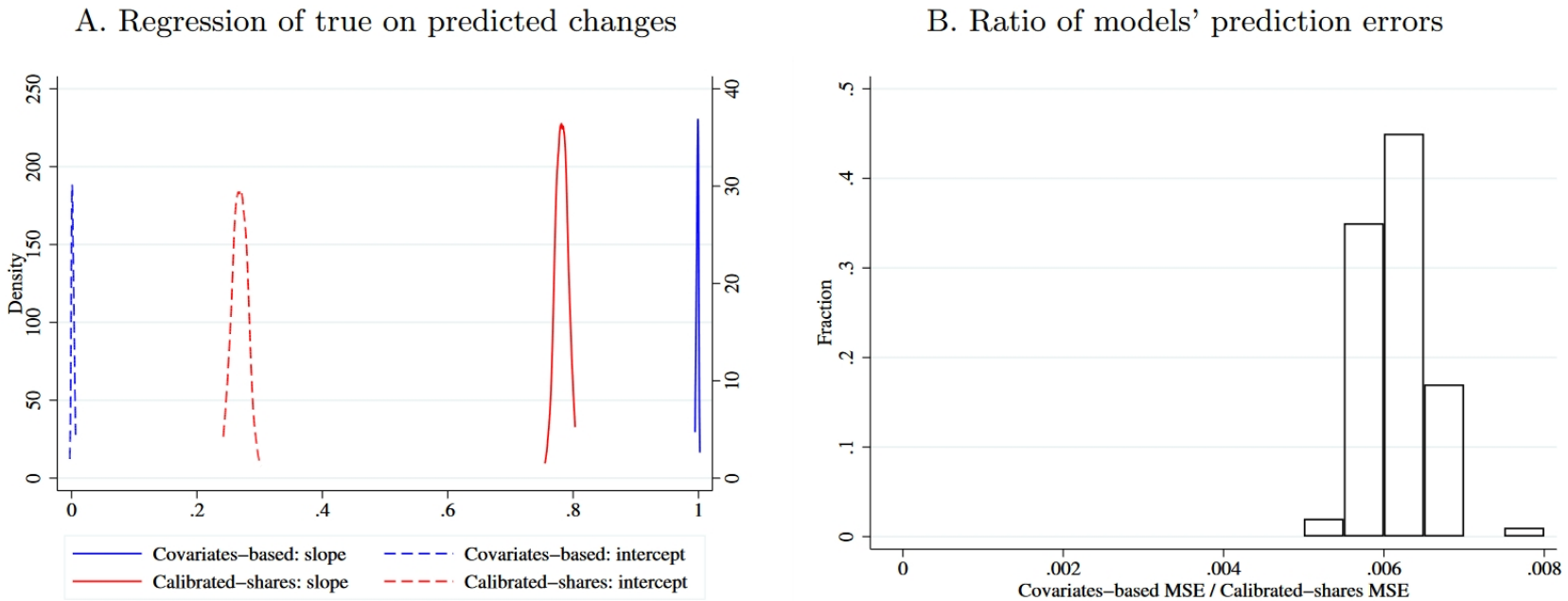
模拟流程:



蒙特卡洛模拟: 两种思路的比较

- 蓝色线代表“协变量方法”的预测结果，100 次模拟得到的斜率的分布集中在 1 附近，截距的分布集中在 0 附近，表明其预测非常准确，而且波动很小。
- 红色线代表“精确帽代数”的预测结果，其预测即不准确又具有较大波动。

Figure 2: Calibrated-shares procedure overfits in Monte Carlo simulations



蒙特卡洛模拟: 解释

- “协变量方法”仅使用少量参数来模型化通勤成本，从而拟合出一个与观测数据不太一样的人口份额，虽然忽略了不可观测的通勤成本，但观测的人口数据中的噪声不太会影响到基准均衡的校准。
- “精确帽代数”方法完美地匹配了基准情形下观测的人口份额，我们可以将其视为一个极其灵活的参数化模型 (相当于为每个区块对的通勤成本都设定了一个待估参数的“饱和模型”)。
- 在噪声很大的时候，后者非常准确地拟合了基准情形 (“训练集”)，但其实是把很多噪声也当作了规律，导致在反事实预测 (样本外预测) 时表现糟糕，这就像机器学习中常说的“过拟合问题”。

蒙特卡洛模拟: 解释

- 如果我们增加生成的模拟样本的数量，比如将人数从 250 万不断增加，可以发现“精确帽代数”的预测结果会不断优化。当每个区块对的人数平均超过 50 个人之后，“精确帽代数”方法的准确性和波动性就会变得非常优秀。
- 这正反映了颗粒状设定下，“精确帽代数”方法的问题来源。

Table 2: Calibrated-shares procedure's finite-sample performance

A. Regressand is continuum change in commuters

I	2.5	5	12.5	25	50	125	250	2560
Calibrated-shares: slope	0.782	0.876	0.948	0.974	0.986	0.995	0.997	1.000
Calibrated-shares: intercept	0.269	0.153	0.064	0.032	0.017	0.007	0.004	0.000
Calibrated-shares: MSE	0.225	0.113	0.045	0.023	0.011	0.005	0.002	0.000

B. Regressand is finite-sample change in commuters

I	2.5	5	12.5	25	50	125	250	2560
Calibrated-shares: slope	-0.408	0.194	0.669	0.835	0.913	0.968	0.982	0.998
Calibrated-shares: intercept	1.724	0.982	0.404	0.202	0.106	0.040	0.022	0.002
Calibrated-shares: MSE	17.022	8.486	3.400	1.699	0.851	0.340	0.169	0.017

蒙特卡洛模拟: 解释

- 我们进一步在生成模拟数据时加入不可观测的通勤成本:

$$\ln \delta_{kn} = \ln \bar{\delta}_{kn} + \ln \lambda_{kn}$$

$$\ln \lambda_{kn} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda \times \text{Var}[\ln \bar{\delta}_{kn}]), \quad \Lambda \in \mathbb{R}^{+}$$

蒙特卡洛模拟: 解释

- 如果存在不可观测的通勤成本，颗粒状设定下，“精确帽代数”的估计依然有偏，“协变量方法”无偏、但波动性增加。
- 在大样本下，“精确帽代数”更加灵活的优势就表现出来了，其预测效果会胜过“协变量方法”。

Table A.2: Monte Carlo simulations with $\lambda_{kn} \neq 1$: Regressand is continuum change

		Slope (mean)		MSE (mean)	
Λ	I	Covariates-based	Calibrated-shares	Covariates-based	Calibrated-shares
0	2.5	0.9985	0.7817	0.0014	0.2252
0	5	0.9992	0.8759	0.0007	0.1130
0	12.5	0.9995	0.9479	0.0003	0.0452
0	25	0.9998	0.9737	0.0001	0.0227
0	50	0.9999	0.9864	0.0001	0.0112
0	125	1.0000	0.9946	0.0000	0.0045
0	250	1.0000	0.9971	0.0000	0.0023
0	2560	1.0000	0.9997	0.0000	0.0002
1	2.5	0.9954	0.9688	6.3762	0.2176
1	5	0.9965	0.9837	6.3749	0.1092
1	12.5	0.9972	0.9933	6.3745	0.0441
1	25	0.9969	0.9965	6.3750	0.0218
1	50	0.9971	0.9982	6.3748	0.0109
1	125	0.9971	0.9994	6.3747	0.0044
1	250	0.9972	0.9995	6.3746	0.0022
1	2560	0.9972	0.9998	6.3746	0.0002

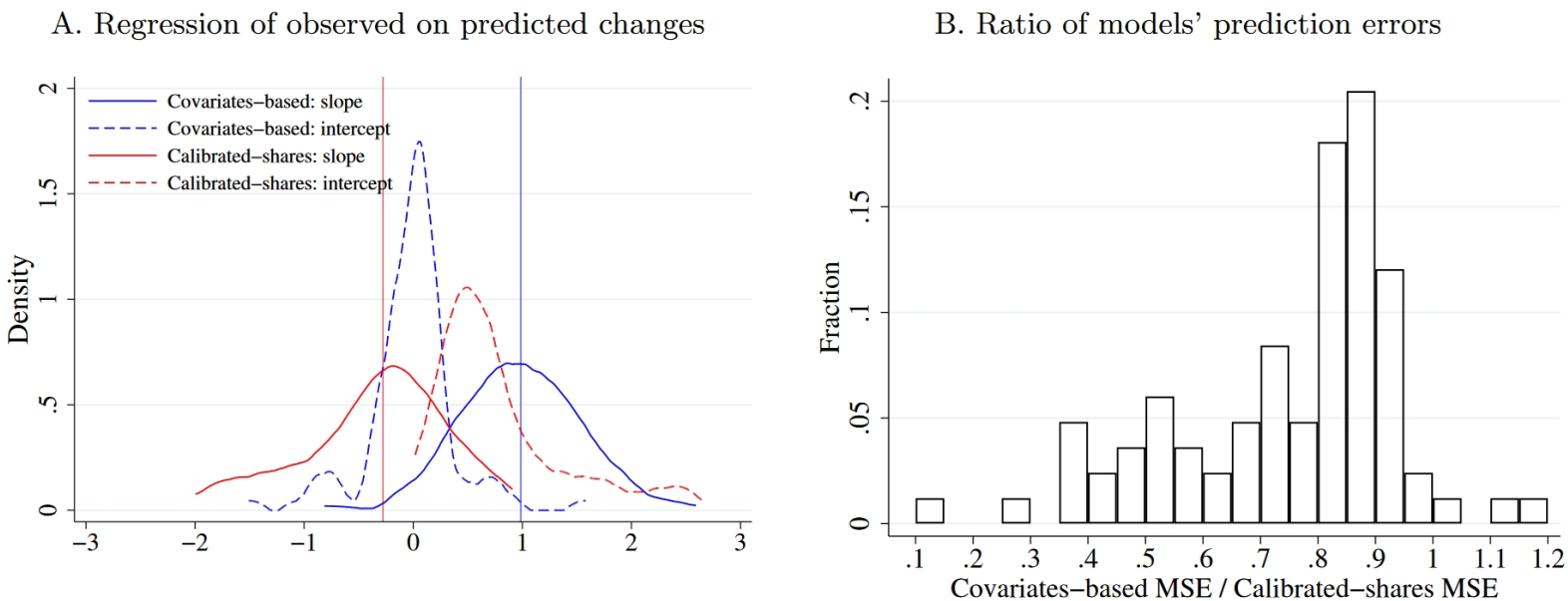
现实中的冲击

- 作者挑选出了 83 个区块作为“处理组区块”, 这些区块具有较大的工人数量, 并在 2010–2012 年雇员数量有较为明显的增长, 作者认为, 这些增长主要是由 workplace-specific 的 shock 导致的, 因此可以将其 model 成这些区块出现了生产率的外生增长。
- 作者首先借助反事实均衡, 在理论上反推出需要多大的生产率变化才能使其实现于现实一致的雇员数量增长。
- 随后, 利用与前面蒙特卡洛模拟一样的模拟流程, 分别用“协变量方法”和“精确帽代数”预测这些生产率的外生增长带来的经济后果。

两种思路的比较

- “协变量方法”(蓝色线) 得到的斜率分布在 1 附近，截距分布在 0 附近，表明其预测效果较为理想。
- “精确帽代数”(红色线) 的偏误依然很大。

Figure 3: Comparison of models' predictive performance across 83 events



一些针对“精确帽代数”的调整方法 (略过)

- 用多个事前期的数据来校准基准均衡。
- 加总到更大一点的地理尺度 (the Neighborhood Tabulation Area, NTA)。
- 使用秩限制奇异值分解 (rank-restricted singular value decomposition, SVD) 方法构建一个近似的人口份额矩阵。

关于两种校准思路的总结

- “精确帽代数”: 使用观测到的数百万个人口份额来校准数百万个通勤成本，会产生严重的过拟合问题。问题的关键在于，观测到的人口份额不太可能是模型中大样本理论作用下的份额。
  - 所以问题关键在于空间尺度的大小？准确说，是观测人口数相对于地区对的数量大小 (最好至少 >50)。
- 一些调整方法能缓解“精确帽代数”在 granular 设定下的不足，合并数据通常帮助不大，但奇异值分解得到的近似矩阵能够很大程度上修正有偏误的估计结果。
- “协变量方法”放弃了观测的人口份额，通过建模出一个有些许偏差的通勤成本，来估计大样本下的人口份额。其反事实估计与理论结果非常接近。
  - 完美无缺？不是！因为只建模了一部分的通勤成本，所以如果在大空间尺度下，可能明显低估通勤/流动成本。
  - “协变量方法”拟合了一个没有 idiosyncratic preference 的经济体，但... 我们关心的真的是这个吗？

目录

有限个体模型和传统连续模型的区别

- 我们能不能将个体的 idiosyncratic preference 引入到空间模型中？
- 目标: 如何考察个体特异性偏好对反事实结果产生的不确定性
- 时序设定: 1) 个体首先选择居住地、通勤地 2) 随后市场出清
  - 先有鸡还是先有蛋: 连续模型 vs 有限个体模型
  - 有限个体模型的均衡价格并不一定收敛到连续模型下的均衡
- 模型思路
  - 从 $\boldsymbol{\nu}^I$ 分布中抽取个体特异性偏好 ν_{kn}^i
  - 均衡价格和均衡数量变为一个分布: workers, wages and prices
  - 反事实分析中，将外生变量从 $\Upsilon \to \Upsilon'$，反事实变化也会变为一个分布
- 有限个体模型下，个人的决策会影响整体的工资和租金

基本设定

- 基本设定上，有限个体模型较连续模型的主要变化：
  1. **有限假定**：假设经济中共有 I 单位工人，每个人可以提供 L/I 单位劳动投入，于是总的劳动投入量为 L 单位
  2. **时序假定**：假定工人首先进行居住地与工作地的决策，实现空间决策之后，各地市场出清再决定最终的价格 (最终价格可能与决策时预期到的不一样，但即使如此，工人也不再调整决策了)。

基本设定

Assumptions about information and expectations

- 经济体外生变量 Economic primitives $\Upsilon \equiv \{L, \{A_n\}, \{T_k\}, \{\bar{\delta}_{kn}\}, \{\lambda_{kn}\}, \alpha, \epsilon, \sigma\}$: 假定所有个体都知道这些信息
- 预期价格 Belief vector $\{\tilde{w}_n\}$ and $\{\tilde{r}_k\}$: 假定所有个体心中对工资和租金的预期价格都是一样的 (point-mass beliefs/common expectations)，这些信念是具有相同经济外生变量的连续体模型中的均衡价格
- 时序设定: 1) 个体首先选择居住地、通勤地 2) 随后市场出清
  - 个体基于预期价格 $\{\tilde{w}_n\}$ and $\{\tilde{r}_k\}$，确定他们效用最大化时的居住地、工作地区位选择 kn
  - 随后，个体不再移动，kn 保持不变 → 个体分布 $\{l_{kn}\}$ 确定
  - 随后，商品、土地、劳动力市场出清，得到一组实现的均衡价格 w_n and r_k

基本设定

- 假设工人知晓基准变量 Υ 并形成了对价格相同的“连续情形下的理性预期”，我们将预期的价格写作 $\{\tilde{w}_n\}$ 和 $\{\tilde{r}_k\}$，同时工人知道自己的特异性偏好 $\{\nu^i_{kn}\}$。与连续模型一致，工人 i 选择在地区 k 居住、在地区 n 工作的间接效用函数为：

$$\tilde{U}^i_{kn} = \epsilon \ln\left(\frac{\tilde{w}_n}{\tilde{P}^{1-\alpha}\tilde{r}_k{}^{\alpha}\delta_{kn}}\right) + \nu^i_{kn} \tag{8}$$

- 从表达式上看，和连续模型唯一的差别，就是价格都变成了预期价格而已：

$$\tilde{P} = \left[\sum_n (\tilde{w}_n/A_n)^{1-\sigma}\right]^{1/(1-\sigma)}$$

有限个体模型的空间均衡

- 由于模型假设了个体先根据预期价格做出居住与工作决策，再由市场出清决定价格，因此需要重新表述一下关于模型均衡的定义。在新的模型中，人口分布 $\{l_{kn}\}$ 定义了有限个体下的通勤均衡 (commuting equilibrium with finite many individuals)，给定 $\{l_{kn}\}$ 之后，再由价格 $\{w_n\}$ 和 $\{r_k\}$ 定义贸易均衡 (trade equilibrium)。

Definition (贸易均衡, Trade equilibrium)

给定人口分布 $\{l_{kn}\}$ 以及基准变量 Υ，所谓“贸易均衡”是指能够使的市场出清条件式 (3) 与式 (4) 成立的工资向量 $\{w_n\}$ 与土地价格 $\{r_k\}$ 向量。

$$A_n \sum_k \frac{l_{kn}}{\bar{\delta}_{kn}} = \frac{(w_n/A_n)^\sigma}{P^{1-\sigma}} Y, \quad \forall n$$

$$T_k = \frac{\alpha}{r_k} \sum_n \frac{w_n l_{kn}}{\bar{\delta}_{kn}}, \quad \forall k$$

有限个体模型的空间均衡

Definition (有限个体下的通勤均衡, Commuting equilibrium with finite individuals)

给定人口总数 I 以及基准变量 Υ、个体特异性偏好的一组实现值 $\boldsymbol{\nu}^I$ 以及工人对各地价格的相同预期 $(\{\tilde{w}_n\},\{\tilde{r}_k\})$，所谓“有限个体下的通勤均衡”是指能够满足以下条件的人口分布 $\{l_{kn}\}$、工资向量 $\{w_n\}$ 和土地价格向量 $\{r_k\}$：

- 工人根据效用最大化选择居住地与工作地：

$$l_{kn} = \frac{L}{I}\sum_{i=1}^{I}\mathbf{1}\left\{\tilde{U}_{kn}+\nu_{kn}^{i} > \tilde{U}_{k'n'}+\nu_{k'n'}^{i}, \forall (k',n') \neq (k,n)\right\}$$

- 给定人口分布 $\{l_{kn}\}$，求得工资向量 $\{w_n\}$ 与土地价格向量 $\{r_k\}$ 以满足贸易均衡。

Definition (连续情形下的理性预期价格, Rational expectations for the continuum case)

给定基准变量 Υ，所谓“连续情形下的理性预期价格”是指与连续模型在同样的基准变量 Υ 下得到的均衡价格 $(\{w_n\},\{r_k\})$ 相等的一组价格向量 $(\{\tilde{w}_n\},\{\tilde{r}_k\})$。

校准基准变量

- 有限个体模型的基准均衡同样需要校准，仅需将价格换成预期价格，得到下面式 (9) 这一对数似然函数：

$$\begin{aligned}\mathscr{L} &\equiv \sum_{k,n} l_{kn} \ln\left[\Pr\left\{\tilde{U}^i_{kn} > \tilde{U}^i_{k'n'}, \forall k'n' \neq kn\right\}\right] \\ &= \sum_{k,n} l_{kn} \ln\left[\frac{\tilde{w}_n{}^{\epsilon}\left(\tilde{r}_k{}^{\alpha}\bar{\delta}_{kn}\right)^{-\epsilon}}{\sum_{k',n'}\tilde{w}_{n'}{}^{\epsilon}\left(\tilde{r}_{k'}{}^{\alpha}\bar{\delta}_{k'n'}\right)^{-\epsilon}}\right]\end{aligned} \tag{9}$$

- 尽管对数似然函数中存在预期价格，但在实际估计中只需加入居住地与工作地的固定效应即可。
- 因此，在校准基准均衡的操作上，可以说，有限个体模型与“协变量方法” (CBA) 是完全一致的，$\epsilon, l_{kn}, T_k, A_n$ 的估计量也一样。

工人事后会后悔吗？

- 有限个体模型建立在一个很重要的时序假设上：工人先按预期价格进行决策和行动，然后再形成均衡价格，即使均衡价格与其预期价格不一致，工人也不再改变决策。
- 可以通过模拟考察均衡价格形成以后工人后悔的比例，来检验这一假设的合理性。
- 作者构建了一个“后悔率 (ex post regret)”指标 χ_i，并通过多次模拟来估计这个指标的大小。具体来说，假设均衡下价格为 $\{w_n, r_k\}$，则一个选择在地区 k 居住、在地区 n 工作的工人 i 的后悔率 χ_i 满足：

$$\max_{k',n'}\left(\epsilon \ln\left(\frac{w_{n'}}{P^{1-\alpha} r_{k'}^{\alpha} \delta_{k'n'}}\right) + \nu_{k'n'}^{i}\right) = \epsilon \ln\left(\frac{(1+\chi_i) w_n}{P^{1-\alpha} r_k^{\alpha} \delta_{kn}}\right) + \nu_{kn}^{i}$$

如果 $\chi_i = 0$，代表工人没有后悔；如果 $\neq 0$，则其大小代表根据预期做出的“真实选择”与事后的“正确选择”的偏差程度。

工人事后会后悔吗？

- 作者生成了 250 万个工人 (本质上是 250 万组 $\{\{v_{kn}^i\}_{k,n}\}_i$) 进行模拟，计算这 250 万个工人各自的后悔率 χ_i。上述模拟一共进行了 10 次 ($s = 1, \cdots, 10$)。
- $\chi_i > 0$ 的比例大约为 4.44%，即超过 95% 的工人不会在事后后悔。
- 在后悔的个体中，后悔率的中位数大约是 0.72%，表明对于后悔的工人来说，其“真实选择”与“正确选择”的偏离大约为 0.72%，并不算大。

Table E.1: Distribution of ex post regrets

		Unconditional distribution					Conditional distribution	
s	Share with regret	p95	p96	p97	p98	p99	Mean	Median
1	0.0442	0.0000	0.0011	0.0042	0.0082	0.0150	0.0106	0.0073
2	0.0433	0.0000	0.0009	0.0039	0.0078	0.0143	0.0102	0.0071
3	0.0446	0.0000	0.0012	0.0043	0.0083	0.0150	0.0106	0.0072
4	0.0446	0.0000	0.0012	0.0043	0.0084	0.0152	0.0106	0.0073
5	0.0437	0.0000	0.0010	0.0040	0.0079	0.0144	0.0103	0.0071
6	0.0444	0.0000	0.0012	0.0042	0.0083	0.0150	0.0107	0.0073
7	0.0447	0.0000	0.0013	0.0043	0.0083	0.0150	0.0105	0.0072
8	0.0445	0.0000	0.0012	0.0043	0.0084	0.0150	0.0106	0.0073
9	0.0452	0.0000	0.0014	0.0045	0.0086	0.0154	0.0109	0.0074
10	0.0444	0.0000	0.0011	0.0042	0.0082	0.0148	0.0106	0.0072
mean	0.0444	0.0000	0.0012	0.0042	0.0083	0.0149	0.0106	0.0072

有限个体模型和传统连续模型的区别

- 将个体的 idiosyncratic preference 引入到空间模型中
- 个体特异性偏好会影响均衡结果和反事实结果
- 时序设定: 1) 个体首先选择居住地、通勤地 2) 随后市场出清
  - 先有鸡还是先有蛋: 连续模型 vs 有限个体模型
  - 有限个体模型的均衡价格并不一定收敛到连续模型下的均衡
- 模型思路
  - 从 $\boldsymbol{\nu}^I$ 分布中抽取个体特异性偏好 ν_{kn}^i
  - 均衡价格和均衡数量变为一个分布: workers, wages and prices
  - 反事实分析中, 将外生变量从 $\Upsilon \to \Upsilon'$, 反事实变化也会变为一个分布

有限个体的模型和传统的连续模型的区别

- 形式上：
  - 传统的连续模型使用大样本理论消除了个体的特异性偏好，因此求得的反事实预测是一个确定的数。
  - 在有限个体模型中，个体特异性偏好的影响并没有被完全消除 (wash out)，因此进行反事实分析时，我们关心的变量的变化应当是一个分布。
- 操作上：
  - 传统的连续模型在校准出基准均衡后，将变量的外生变化一同代入反事实均衡条件式 (5)–(7)，即可求得内生变量的相对变化。
  - 有限个体模型使用与“协变量方法”相同的方式校准出基准均衡之后，基于校准出来的基准均衡，生成一个与现实人数相同的有限个体样本，并基于一个给定的分布，为个体逐一生成特异性偏好。随后，纳入外生冲击，计算出每个个体在外生冲击下的“通勤均衡”，进而计算出“贸易均衡”，此时完成一次模拟。
  - 通过多次模拟 (即反复生成样本)，就可以得到对关心的变量的反事实变动的一个预测分布，将这个分布的均值作为有限个体模型下的反事实预测。

目录

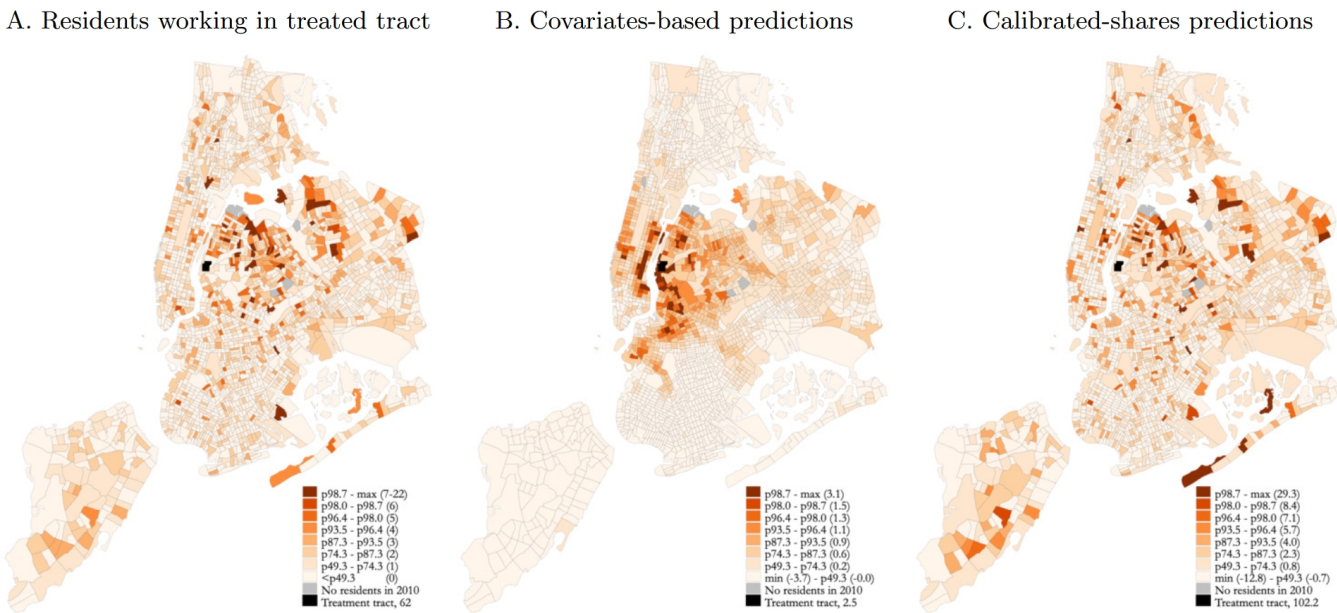
亚马逊 HQ2 的冲击

- 下面作者用连续模型的两种思路与有限个体模型分别估计一个现实案例的经济影响: 亚马逊计划在长岛市 (Long Island City) 建立第二总部 (HQ2)。
- 背景:
  - 2017 年, 亚马逊宣布了建立 HQ2 的计划, 他们将在美国寻找一个新的总部地点, 以便扩大业务并提供更多就业机会, 长岛市成为了亚马逊 HQ2 的候选地之一。
  - 2018 年 11 月, 亚马逊宣布在长岛市建立 400 万平凡英尺的办公区域、雇佣超过 25,000 位雇员。
  - 2019 年 2 月, 这个计划遭受了当地势力的剧烈反对而被迫流产。当地的政治家和社区居民一方面担心政府对亚马逊的补贴, 另一方面担心亚马逊的计划会导致当地缙绅化 (gentrification)。
- 作者将建立 HQ2 的外生冲击刻画为长岛市所在区块 (“处理组区块”, 记为 n^*) 生产率的外生增长, 因此作者首先分别用“协变量方法”和“精确帽代数”反推出使得长岛市所在区块增加 25,000 个工人对应的生产率增长规模, 分别记为 $\hat{A}_{n^*}^{\text{CBM}}$ 和 $\hat{A}_{n^*}^{\text{CSP}}$。
- 接下来, 用三种思路分别估计这一生产率的外生变化带来的影响。

三种估计思路的对比

- Figure 6 是纽约市在处理组区块工作的工人的居住地分布及其反事实变化，其中黑色的区块代表长岛市，颜色越红代表在住的工人越多。
- 图 A 是 2010 年的观测数据：HQ2 建立之前，在处理组区块工作的工人大多居住在处理组区块的东边。

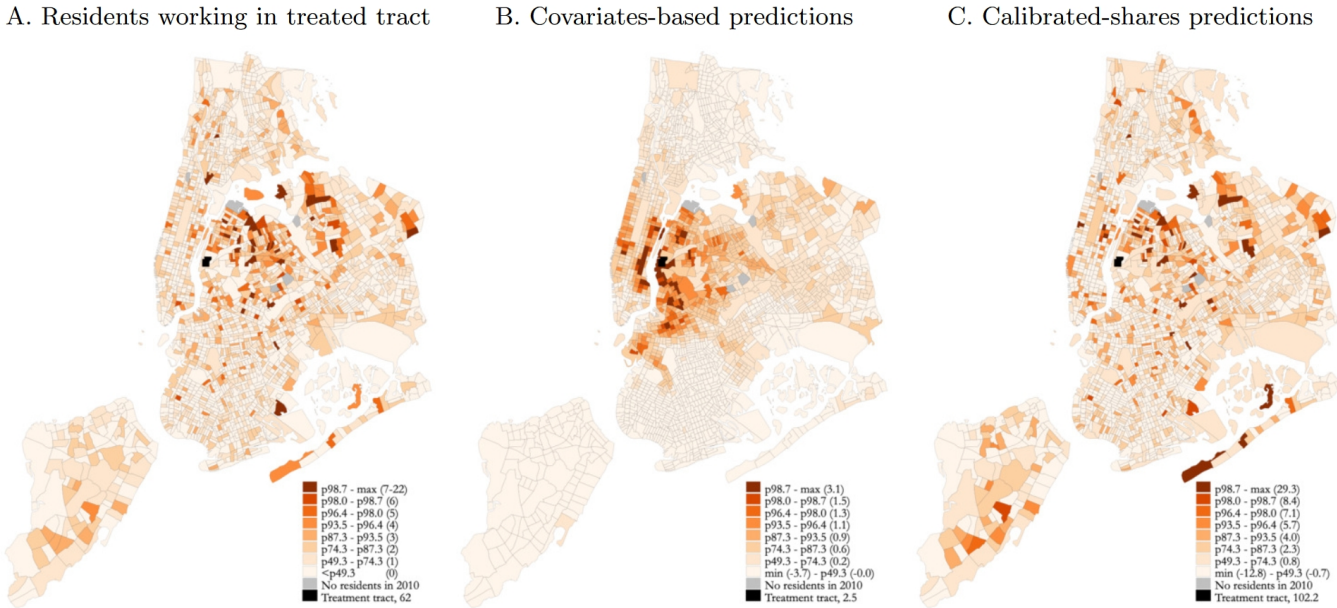
Figure 6: Amazon HQ2 counterfactual change in residents



三种估计思路的对比

- 图 B 和图 C 分别提供了“协变量方法”与“精确帽代数”所预测的居住分布的反事实变化。前者表明若 HQ2 建立，在处理组区块工作的居民将集聚到该区块周围居住；后者的估计结果则表明，居住人口的增加地更多的地区 (图 C) 基本上就是原来人口分布地多的地区 (图 A)。
- 根据“协变量方法”，处理组区块的居住人口大约增加 2.5%；根据“精确帽代数”，处理组区块的居住人口将显著增加约 102.2%(有点夸张的估计结果)。

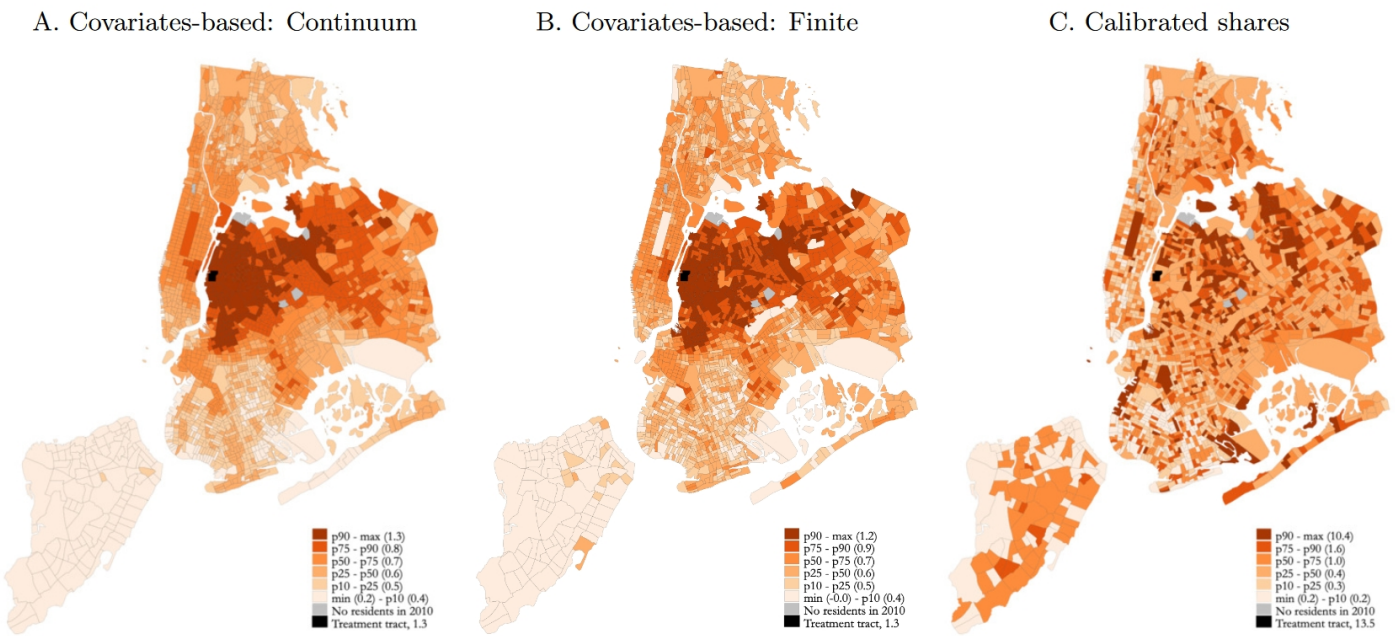
Figure 6: Amazon HQ2 counterfactual change in residents



三种估计思路的对比

- Figure 7 展示了土地价格的反事实变化。在“精确帽代数”的估计下，实际租金的变化 $\hat{r}_k/\hat{P}$ 非常明显，增长最大的区块大约增长 10.4%。
- “协变量方法”(图 A 代表连续模型，图 B 代表有限个体模型) 预测的价格变化较小，增长最大的区块大约增长 1.3%(图 A) 或 1.2%(图 B)，连续模型和有限个体模型的预测非常相似。

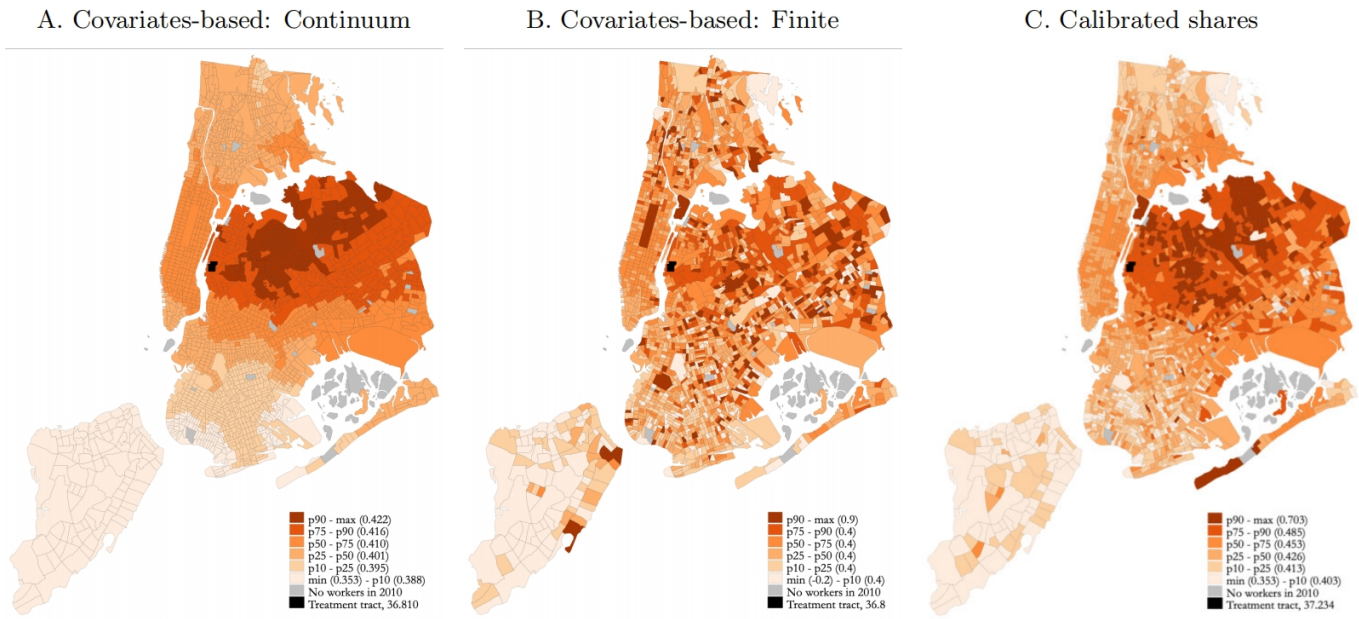
Figure 7: Predicted changes in rents



三种估计思路的对比

- 如果考察各区块工资的变化，则三种思路的预测较为接近。

Figure F.2: Predicted changes in wages



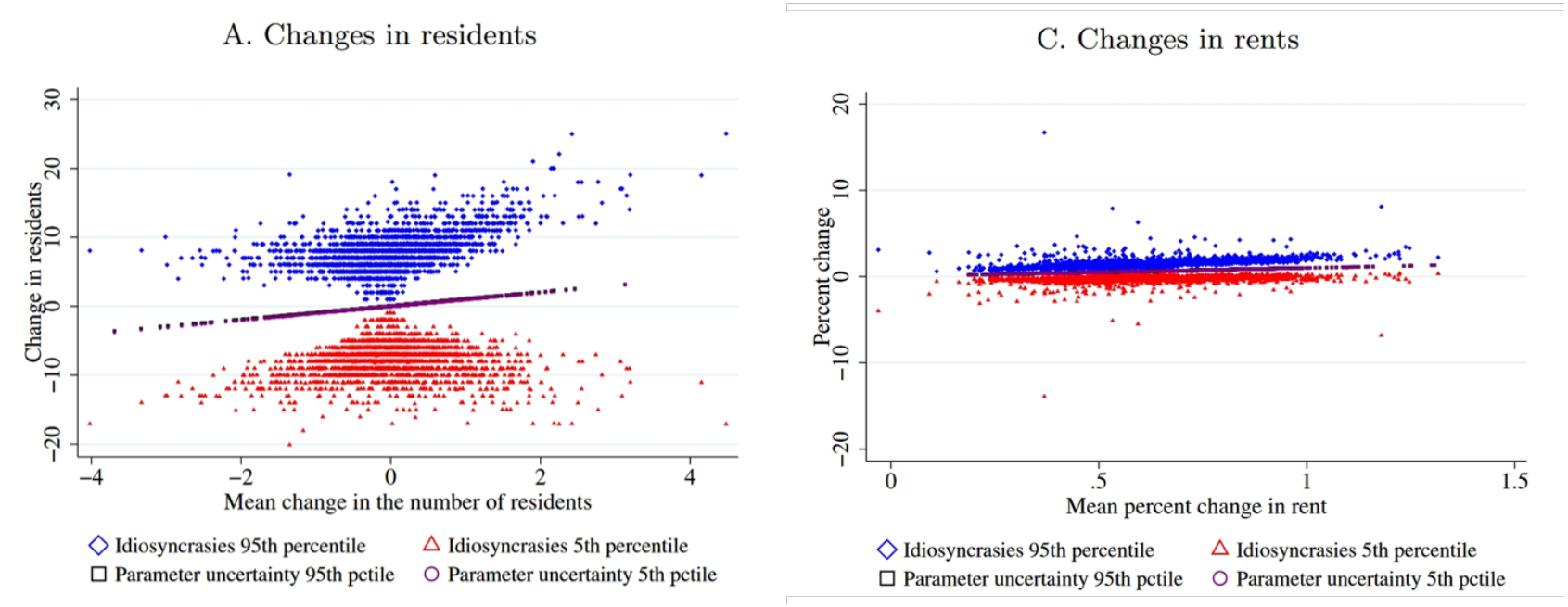
反事实变化的“统计推断”

- 有限个体模型可以得到反事实变化的分布，而不仅仅是一个“点估计值”: 作者构建了 100 个包含 250 万个个体的样本 (即在校准出来的基准均衡的基础上进行 100 次模拟)，并分别计算其通勤均衡与贸易均衡，从而得到 100 个反事实预测，这 100 个预测就构成了一个预测分布
- 不确定性的来源: 参数校准的不确定性 or 个体特异性偏好的不确定性
- Figure 8 分别考察了各区块的居民数量、工人数量、土地价格与工资的变化，蓝色和红色代表由个体特异性偏好导致的不确定性，蓝色代表各区块反事实变化分布的 95% 分位数，红色代表 5% 分位数 (从而得到 90% 置信区间)

反事实变化的“统计推断”

- 图 A 是各区块居民数量的变化，各区块 90% 置信区间均包含 0，因此可以说，在统计意义上各区块居民数量的反事实变化是无异于 0 的。
- 图 C 也类似，只有大约 15% 的区块的 90% 置信区间是不含 0 的。

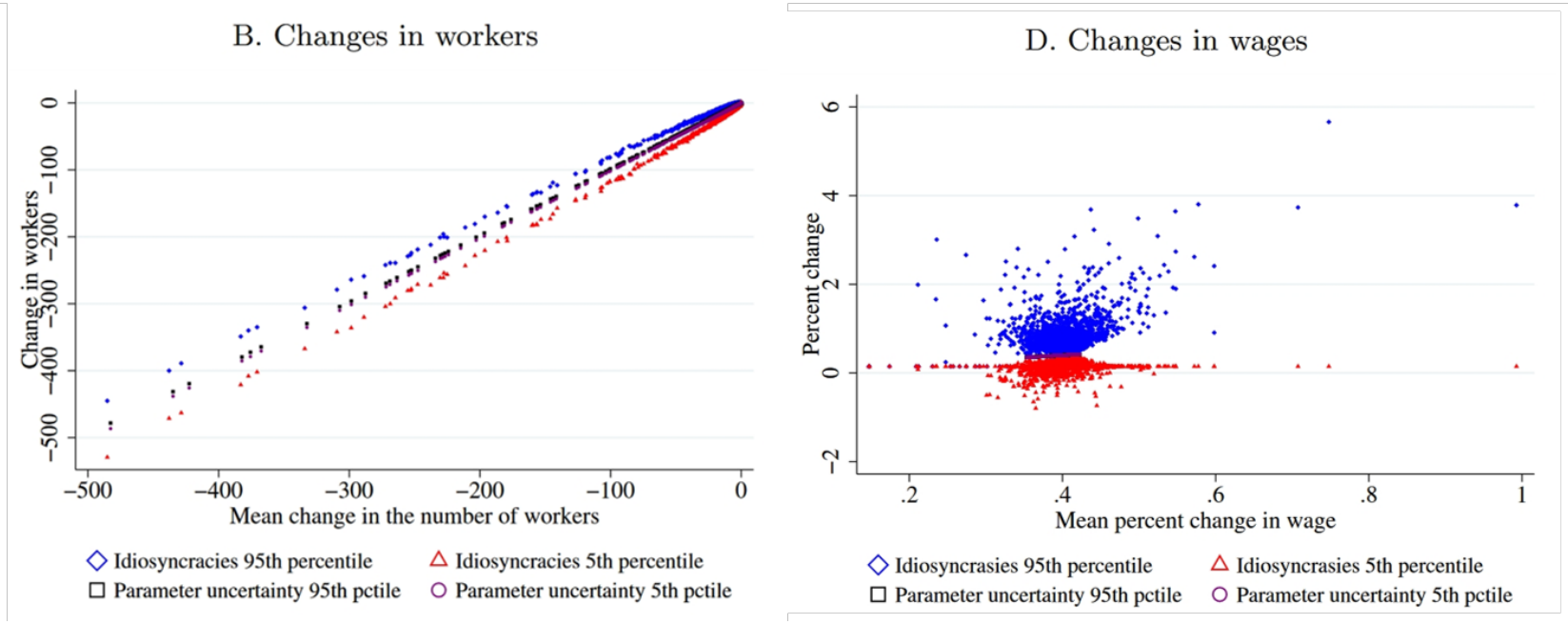
Figure 8: Variation in counterfactual changes: Idiosyncrasies vs. parameter uncertainty



反事实变化的“统计推断”

- 图 B 是各区块工人数量的变化，对于小部分地区而言，其工人数量变化之大已经明显超过了不确定性，其 90% 置信区间不包含 0，实际上，大部分地区 (上百个区块) 的工人数量变化是不大的，其置信区间依然无异于 0。
- 图 D 是工资变化的分布，依然是大部分地区的 90% 置信区间包含 0。

Figure 8: Variation in counterfactual changes: Idiosyncrasies vs. parameter uncertainty



目录

总结

- 精细的空间数据拓展了可研究的领域，但如果直接使用传统的连续模型，可能会有一些问题。
- 蒙特卡洛模拟的结果表明，在颗粒状设定下，“协变量方法”的预测效果往往更好一些，常用的“精确帽代数”往往是有偏的。
- 我们可以对经典模型做出一些调整，构建一个有限个体的空间模型，在这个模型中，个体的异质性没有被足够多的样本消去，使得变量的反事实变化并非一个确定的数，而是一个随机变量。
- 在估计时，我们可以人为构建出与现实类似的有限样本情形，借助这一构建的样本来计算变量的反事实变化；重复这一过程，可以估计出变量反事实变化的分布，这就为统计推断提供了可能。
- 在亚马逊 HQ2 的例子中，我们看到了经典连续模型估计出的不为 0 的反事实变化，在考虑个体异质性带来的不确定性时，很可能在统计意义上其实是无异于 0 的。我们需要重新审视一些以往的研究了。

感谢倾听!

希望对大家有所帮助!