Big Data in Spatial Economics

Lecture 1: Overview of Current and Future Research

Victor Couture (UC Berkeley) and Allison Green (Princeton)

Prepared for PhD289 at UC Berkeley and ECON720 at Yale University.

Fall 2019

Lecture Plan

- Lecture 1: Overview of Current and Future Research
  1. Review of new data sources
  2. A detailed example: Smartphone GPS data
  3. How to find a big data research question?
  4. Ethics of big data research

- Lecture 2: Working with Big Data
  1. Basic principles
  2. `Stata` for big data
  3. Alternatives to `Stata` for really big data (`SQL`, `Python`, `Julia`)
  4. Machine learning for high dimensional/complex data

Lecture 1: Objectives and Plan

- New data sources in urban and spatial economics:
  1. Nielsen scanner data
  2. Credit card transaction data
  3. Full banking record data
  4. Consumer Credit Panel-Equifax data
  5. Internal Revenue Service data
  6. Cell phone record data
  7. Satellite and image recognition
- Spatial data from technology firms
  - How to start a collaboration?
  - Overwiew of existing spatial research (Google, Facebook, Alibaba, etc.)
- Using Smartphone GPS data as an example:
  - How to assess data quality? How to find a research question?

Nielsen Scanner Data: Overview

- Not so new anymore, oldest of "new" spatial data source.
  - Dozens of top papers since 2015!
- Two main datasets:
  1. Consumer panel data
     - Information on each product bought by a panel of users (mostly in grocery stores) at UPC level.
  2. Store level data
     - Weekly prices and sales volumes at UPC level for each store.
- Access: Relatively easy through Kilts Marketing Center at Chicago Booth.

Nielsen Scanner Data: Key Spatial Papers

- *Retail Globalization and Household Welfare: Evidence from Mexico*, Atkin, Faber, Gonzalez Navarro, JPE, 2018
  - Use consumer level and store level Nielsen data from Mexico.
  - Foreign store entry: reduce price in competing stores and capture large market share.
  - Identification: Store entry is an event study in space.

- *Food desert and the cause of nutritional inequality*, Allcott, Diamond, Dube, Handbury, Rahkovsky, Schnell, QJE, 2019
  - Use consumer level and store level Nielsen data from US.
  - Find food availability in space doesn't contribute much to nutritional inequality.
  - Identification: Supermarket entry and household move.

Credit Card Transaction Data: Overview

- New data source in spatial economics, first papers yet unpublished.
- Information available varies, ideally location of transaction, amount, and type of goods.
- Access
  - Visa actively collaborated with Stanford e-commerce paper team *(next slide)*
  - Visa was actively talking to other business schools.

Credit Card Transaction: Key Spatial Papers

- *The Geography of Consumption*. WP. 2019. Agarwal, Jensen, Monte.
  - Sample from 2003: 1.7M transaction, 70K consumers.
  - First spatial use of credit card data, relatively coarse geography and product type.
  - Estimate gravity in consumption expenditure.

- *Assessing the gains from e-commerce*. WP. 2019. Dolfen, Einav, Kleinow, Klopack, Levin, Levin
  - Collaboration with co-authors from Visa
  - Universe of both Visa credit and debit transaction, card ID, merchant ID, zip code geography (street address in later years.)
  - Estimate value of e-commerce from willingness to travel i.e., how people substitute online to offline expenditure as distance to offline store varies.

Banking Data: Overview

- New data source in spatial economics, first papers yet unpublished.
- Information from a single bank about transactions and other account information.
- Access:
  - JP Morgan data seems available if you work for them, do-able for a PhD student.

Banking Data: Key Spatial Papers

- *Is Online Retail Killing Coffee Shops? Estimating the Winners and Losers of Online Retail using Customer Transaction Microdata*. WP. 2019. Relihan
  - JP Morgan Banking data
  - Finds people visit coffee shop more after becoming online grocery shoppers (exploits granularity and scope of data.)

- *The Geography of Consumption Inequality*. No papers yet. Diamond and Moretti.
  - Look at how what people buy and how much they save in space.
  - Data better than CEX to measure savings.
  - Puts consumption inequality in relation with spatial variation in price indices.

Internal Revenue Service (IRS): Overview

- One of numerous admistrative record data sources increasingly in use since 2010
  - Data even more detailed in Scandinavian countries
- IRS data has information about incomes and taxes (e.g., 1040 and W-2 forms)
  - Unique social security number allows to track people through space and time.
  - Tax returns allow to link parents to their children
- Access:
  - Through calls for proposal from the Statistic of Income (SOI) division of the IRS, see https://www.irs.gov/statistics/soi-tax-stats-soi-working-papers

Internal Revenue Service (IRS): Key Spatial Papers

- *Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States* QJE 2014, Chetty, Hendren, Kline, and Saez.
  - Finds positive relationship between income of parents and income of their children.
  - Finds large spatial variation in the strength of that relationship, i.e., some regions have more intergenerational mobility than others.

- *Employment Hysteresis from the Great Recession*, JPE 2019, Danny Yagan
  - Finds that areas that experienced larger increase in unemployment during the Great Recession have lower working-age employment even in 2015.
  - Finds that out-migration did little to alleviate these losses.

Credit Consumer Panel-Equifax: Overview

- Data available since 2010, only recently used in urban economics.
- Information on credit and debt, some demographics, and census block of residence for 5% of US population since 1999.
- Access:
  - Accessing the micro-data requires collaborating with a Fed researcher.

CCP-Equifax Data: Key Spatial Papers

- *Gentrification and residential mobility in Philadelphia.* RSUE 2016. Ding, Hwang, and Divringi
  - Exploits panel nature of data to study displacement.
  - Finds that poor residents (i.e., low credit scores) are no more likely to exit gentrifying neighborhoods.

- *The Long-Run Effects of Neighborhood Change on Incumbent Families.* WP 2019, Baum-Snow, Hartley, and Lee.
  - Links children and parents
  - Identify low income children whose neighborhood experienced a positive Bartik shock.
  - Estimates positive outcomes for these children 15 years later.

Cell Phone Data: Overview

- Data around for years, but ambitious ideas on how to use it emerging now.
- Simplest data only tells cell phone towers where a device "pings" .
  - Available in many countries (see Kreindler and Myauchi paper) including very poor ones.
- Richer data with contact network available in Switzerland (Puga et al. 2019) and China (Barwick et al. 2019).

Cell Phone Data: Key Spatial Papers

- *Measuring Commuting and Economic Activity inside Cities with Cell Phone Records*. WP. 2019. Kreindler and Myauchi.
  - Records of nearest cell tower and timestamp for both texts and voice calls in Sri Lanka and Bangladesh
  - Combine call location with quantitative spatial model to infer wages.

- *Information, Mobile Communication, and Referral Effects*. WP. 2019. Barwick, Liu, Patacchini, and Wu
  - Universe of cell phone records for 12 months in large Chinese city.
  - High frequency of phone calls correlate with worker flows, suggesting job referral.

Satellite and Image Recognition: Overview

- First use of satellite data in economics famous paper *Measuring Growth From Outer Space*. AER. 2012. Henderson, Storeygard, Weil.
  - Use night at lights to measure economic activity.
- Review paper: *The View from Above, Application of Satellite Data to Economics*. JEP. 2017. Storeygard and Donaldson.
- Promising new applications based on imagine recognition.
  - e.g., Use Google Street view or Satellite data to measure poverty, car ownership, slums, etc.
- Easy accessibility – sometimes you pay for images.

Satellite and Image Recognition: Newer Spatial Papers

- *Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life*. EI. 2017. Glaeser, Kominers, Luca, Naik.
  - Review potential for big data in urban economics.
  - Discuss their use of Google Street View image to measure poverty, safety, etc.

- *The Welfare Consequences of Formalizing Developing Country Cities: Evidence from the Mumbai Mills Redevelopment*. WP. 2019. Tsivanidis and Gechter.
  - Use deep learning approach to identify slums from satellite image.
  - Start with training sample that you know are slums, and train machine to recognize slum features (e.g., narrow brown rectangles) in pictures.
  - Allows them to see areas changing in/out of slums through time, and measure the impact of redevelopment on slums in Mumbai.

Data from Technology Firms: Overview

- Many tech firms produce geo-coded data available country-wide or even world-wide!
- You offer prestige of collaboration with great university and possibility of media coverage.
- They offer great data and lots of local and practical knowledge and ideas.
- Very active and exciting area of research recently, too many papers to cover all.
  - Review below biased towards my own work and that of my co-authors

Data from Technology Firms: Access

- Attend industry event and seek/accept offers to meet in person.
- Most firms have a research division, support to academics varies.
- Need high rank support within firm.
  - Gain support from two decision makers to reduce risk if one leaves.
  - Seek skilled partners within firm to help put data together.
  - Former or aspiring academics make great partners.
- Your university helps you draft favorable agreement and signs it for you:
  - Helps you retain your academic freedom to publish whatever results you find.
  - Disclosure regulation in, for instance, California is a gray area that adds risk.

Spatial Data from US Technology Firms

Google

- Hard to collaborate with for spatial economists, but some of their data can be scraped.

- *Mobility and Congestion in Urban India*. WP. 2018. Akbar, Couture, Duranton, Storeygard.
  - Simulate million of trips on Google Maps to study mobility and congestion in India .
  - Now expand to whole world including countries with no travel data.

- *Measuring the Value of Urban Density*. WP. 2016. Couture
  - Use data from Google Place to get restaurant location and characteristics.
  - Merge with NHTS travel data to understand willingness to pay for restaurant variety.

Spatial Data from US Technology Firms

Facebook

- Willing to collaborate with economists, but hard to access.
- Has recent backlash made things harder?
- *The Economic Effects of Social Networks: Evidence from the Housing Market*. JPE. 2018. Bailey, Cao, Kuchler, Stroebel.
  - People more likely to buy a house if their friends in other cities experienced rapid house price growth.
  - Identification: location by time fixed-effects help a lot + individual demographics.

Yelp

- Seems willing to collaborate with economists, or else can be scraped.
- *How Segregated is Urban Consumption* JPE. 2019. Davis, Dingel, Monras, Morales.
  - Scrape Yelp reviews to identify where reviewer live and work.
  - Measure consumption segregation within venues.

Spatial Data from US Technology Firms

Airbnb

- Willing to collaborate with economists.
- *Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications*. WP. 2017. Coles, Egesdal, Ellen, Li, Sundarajan.
  - Collaboration with Airbnb People, tract level data on usage, combine with Zillow house prices, a correlation study.

Uber

- Willing to collaborate with economists, completed projects, but I am not yet aware of a spatial paper with Uber data.
- *Is Uber a substitute or complement for public transit*. JUE (2018), Hall, Price, and Palsson.
  - Look at impact of Uber entry on transit use, no data from Uber.

Spatial Data from US Technology Firms

Amazon: Unwilling to share data with academics, as far as I know.

LinkedIn

- Awarded access to researcher on a competitive basis.
- *The Impact of Restricting Labor Mobility on Corporate Investment and Entrepreneurship*. WP. 2018. Jeffers
  - No strong spatial component but shows potential of data.

Infutor

- Tracks address history of (almost?) all US individuals since 1980, plus some demographic information like age and gender.
- Stanford bought that data, Yale in process of buying it.
- *The Effects of Rent Control Expansion on Tenants, Landlords, and Inequality: Evidence from San Francisco*. AER 2019. Diamond, McQuade, and Quian

Spatial Data from Chinese Technology Firms

Alibaba

- High willingness to collaborate with economists.
- Access through Luohan Academy or AliResearch?
- *Connecting the Countryside via E-Commerce: Evidence from China*. AER-I (2019), Couture, Faber, Gu, Liu.
  - Alibaba agreed to randomly enter villages.
  - Access to universe of transaction and shipment data at village level.

Tencent

- So far, low willingness to collaborate with economists.

Data from Chinese Technology Firms

Baidu

- The Google of China, Baidu Maps similar to Google Maps.
- Willing to collaborate with economists on projects of joint interest.
- *Subways and Road Congestion*. WP 2019, Gu, Zhang and Zou.
  - Baidu shared road segment speed data at different point in time (before and after Subway opening) with research team.

Data from Technology Firms outside of the US and China

Europe

- General Data Protection Regulation in in EU might make data access harder.

Latin America

Asia

- Flipkart has almost 50% e-commerce market share in India + mobile payment, etc.
- Not aware of any papers yet with that company.
- Owned by Walmart, which is not currently known to collaborate.

Africa

- Work on network effects due to introduction of mobile money in Kenya (AER 2014, Jack and Sury)

Smartphone Data: Overview

- Very new data source – first papers emerging now
- Significant entry cost due to data size
- Existing papers:
  - *The effect of partisanship and political advertising on close family ties*. Science. 2018. Chen and Rola
  - *Income Growth and the Distributional Effect of Urban Spatial Sorting*. WP. 2019. Couture, Handbury, Gaubert, Hurst.
  - *Experienced Segregation*, Athey, Ferguson, Gentzkow, Schmidt, WP, 2019

Smartphone Data: Couture, Dingel, Green, Handbury

1. Smartphone movement data description.
2. Quality checks required when using a new data source.
3. What to do if you access new data before having a research question.
4. Challenges of working with large datasets (next lecture).

Smartphone Data Description: Overview

- Smartphone movement data from apps' locational services.
- Raw movement data intersected with basemap of polygons (usually buildings) to generate "visits".
- Each visit linked to a unique location, device, and time stamp.
- Visits assigned attribution scores that measure how confidently we know if a visit actually occurred
- 20 billion visits to commercial locations.
- Assign home as primary residential location where device spends nights

Smarphone Data Description in 100 largest CBSAs

- 69 million smartphone devices with assigned home observed during September 2016-present.
- 6 months between first and last observation of a device on average.
- 600,000 identified chain locations
- 3 billion trips to identified chain locations, 300 million start from home.
- 8 trips from home to identified chain locations per month per device, conditional on making at least one such trip.
- 9.7 million devices are movers who change home locations.

Smartphone Data Description: Building-level Demographic Data

- A third party uses surveys, marketing data, and administrative datasets to assign characteristics like race, income, and building size at the address level.
- The smartphone data provider merges these to each residential building in the basemap.
  - 36 percent of devices have building-level characteristics.
  - Most of these buildings are perfectly homogeneous.
- 0.85 correlation between racial demographics reported for Census blockgroups and building-level racial demographics aggregated to the blockgroup level

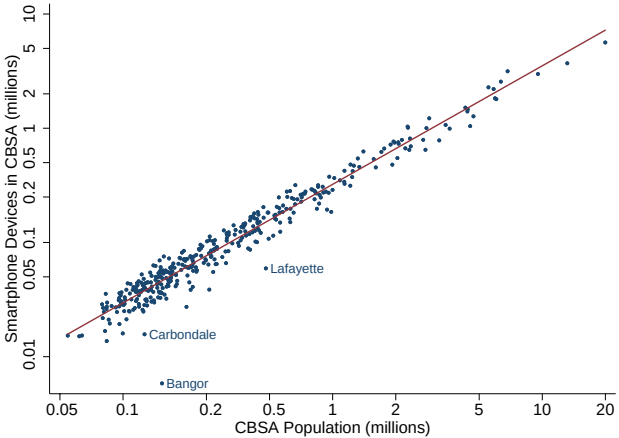
How to assess quality of new data source

1. Sample selection: is data representative of country as a whole?
   - Does geographical distribution of data match that of population?
   - Does income distribution of data match that of population?

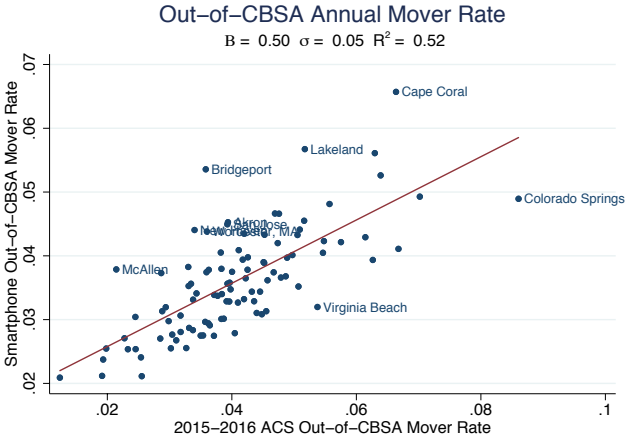
2. Can you match key facts from other data sources?
   - E.g., do smartphone visits resemble those from existing travel surveys?

3. Are new facts from your data sensible?
   - E.g., explore which chains are visited by rich or poor, black or white, etc.

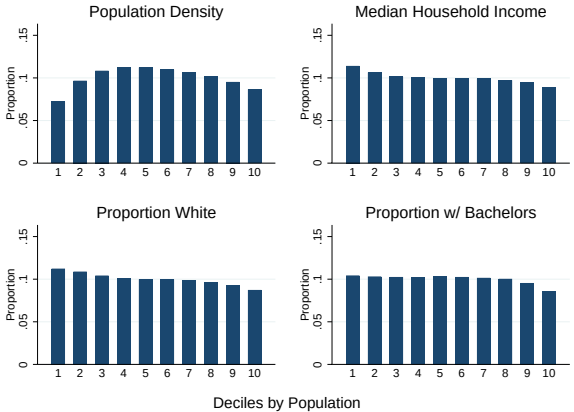
Device counts mirror CBSA populations



Device out-of-CBSA mover rates match ACS data

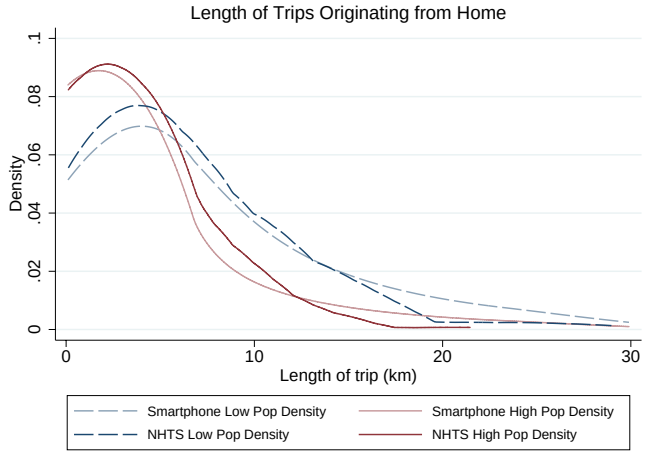


Device residences are representative of US blockgroups



Proportion of devices residing in a blockgroup-characteristic decile is close to 10%.

Device trip lengths comparable to NHTS



Most-visited chains

Category	Chain	Actual Locations	Basemap Establishments	Total Trips
Restaurant	McDonalds	15,828	14,914	260,940,695
Restaurant	Starbucks	14,000+	7,635	60,936,540
Restaurant	Burger King	7,257	7,011	55,056,094
Retail	Walmart	5,358	5,419	60,186,188
Retail	Walgreens	8,100	8,253	40,089,111
Retail	Dollar General	15,015	12,190	21,266,898
Bank	Bank of America	4,600	4,481	34,384,671
Bank	Wells Fargo	6,300+	5,456	28,926,994
Bank	Chase	5,300+	5,572	19,458,021
Gas Station	Shell	14,000	14,266	109,282,462
Gas Station	7Eleven	7,871	7,701	75,860,296
Gym	Planet Fitness	1,500	1,433	12,325,877
Gym	LA Fitness	800+	661	8,068,608
Grocery	Kroger	2,769	2,688	62,003,642
Grocery	Publix	1,231	1,195	19,144,804
Grocery	Safeway	1650+	2,390	10,537,419

Ranking of restaurant chains by high-income visits

Within top 25 chains, rank correlation > 0.9 across block groups, buildings, and marketing surveys:

Restaurant Ranking	High income block groups (census)	High income building	High Income Market Potential Index
1	Chipotle	Chipotle	Panera Bread
2	Panera Bread	Starbucks	Chipotle
3	Starbucks	Panera Bread	Starbucks
4	Dunkin Donuts	Dunkin Donuts	Red Lobster
5	Chick-Fil-A	Jack in the Box	Dunkin Donuts
...			
20	Applebees	Burger King	Dennys
21	Pizza Hut	Sonic	Burger King
22	Burger King	Dominos Pizza	Taco Bell
23	Arbys	Pizza Hut	Arbys
24	Dominos Pizza	Arbys	KFC
25	KFC	KFC	Pizza Hut

Block groups/buildings have high income if median block group/building income $> \$100,000$.

Market Potential Index for earners $> \$100,000$ computed from ESRI data by Couture and Handbury (2017).

Ranking chains by visits by racial demographics

Restaurant ranking	Black homogeneous block groups (census)	Black building in minority black block group	White building in minority black block group
1	KFC	KFC	Arbys
2	Burger King	Papa Johns	Dunkin Donuts
3	Wendys	Chick-Fil-A	Dairy Queen
4	Subway	Burger King	Applebees
5	Dominos Pizza	Wendys	Chick-Fil-A
		...	
20	Chipotle	Dairy Queen	Burger King
21	Panera Bread	Olive Garden	Subway
22	Dennys	Chipotle	KFC
23	Chilis	Panera Bread	IHOP
24	Dairy Queen	Starbucks	Dennys
25	Starbucks	Dunkin Donuts	Jack in the Box

Block group or building is race r if composition is greater than 90%.

- Columns 1 and 2: 0.72 correlation vs Columns 2 and 3 -0.17 correlation.

Ranking chains by visits by racial demographics

How to Find a Research Question with New Data?

- Ideally, come up with research question, then find new data necessary to answer it.
  - My strong preference for almost all my projects.
  - Good way to find to new data sources!
  - E.g., We wanted to write a paper about congestion in developing country and landed on Google Maps.
- Data providers will ask you for a research proposal anyways.
- Smartphone data: We got the data before our research idea.

How to Find a Research Question with New Data?

- Rule: New data must answer research question better than existing data
  - Lots of ideas did not meet that threshold.

- What is the most unique feature of new data?
  - In smartphone case, very high frequency panels with precise geocode
  - We looked into event study in space (Glaeser et al. 2017)
    - E.g., mobility before and after introducing transit stops.
    - Didn't work well; even biggest data get sparse fast.

How to Find a Research Question with New Data?

- Other unique feature of smartphone data: you see visitors to each commercial venue.
  - We landed on measuring preferences for social exposure.
  - Fits well with our expertise as a group of co-authors (not to be neglected).

- Other team using smartphone data – Athey and Gentzkow – also landed on project measuring "experiential segregation".

Some Thoughts on Ethics and Big Data

- New data sources create new ethical questions for research
- Largely up to researchers on the frontier to define ethical best practices (with input from IRB boards)
- Major considerations:
  - Independence from private sector influence.
  - Did people consent to have their data used in this way?
  - Is data provider transparent about methodology and sources?
  - Is data handled securely by people authorized to access it?
  - Are there biases in data or procedures used to process it?
    - e.g. Known racial bias in machine learning for facial recognition.