Method, Selecting Variables, and Collecting Data

HAYDAR KURBAN

DEPARTMENT OF ECONOMICS & CENTER ON RACE AND WEALTH (CRW)

HOWARD UNIVERSITY

HKURBAN@HOWARD.EDU

Strategies for Selecting Appropriate Research Methods and Variables

- **Research Design:** A plan or strategy to carry out research. It is a blueprint of the study.
- **Research Question**: Identifies/describes topics to be studied and used to generate hypotheses
- Research design allows a researcher to develop or select "appropriate methods" and procedures to provide "credible answers" to the research questions and test hypotheses with a "high degree of confidence"
- Selected research methods should yield " robust results" or the strongest possible results
- Appropriate empirical methods yield robust results if data is "right"
- Usually linear regression is chosen as an appropriate method (quantitative method)
- Lack of data yields biased results
- Observational data versus experimental data

A Simple OLS Model: Effect of Treatment (T) on Observed Outcome (Y)

- **$Y_i = \beta_0 + T_i\beta_1 + \varepsilon_i == X_i\beta + \varepsilon_i$ (X=[1 T])**
- dichotomous treatment variable: T=1 if treated, 0 otherwise
- homogeneous treatment effect (β)
- linear
- no covariates
- **Least Square estimate yields β^{OLS} =average outcome Y for T=1 – average outcome for T=0**
- Key assumption of least-squares: E(X'**ε) = 0**
- **That is treatment is uncorrelated with omitted variables**

Four Solutions to this Problem

1. Randomized Controlled Trial

RCT is designed to ensure key OLS assumption: $E(T'\varepsilon)=E(T'W)=0$.

2. .Natural. Experiments

Find similar observations with different treatment for arbitrary.

reasons (e.g. regulatory rules, law changes)

- Difference-in-Difference. Estimates
- Discontinuity design (physical boundaries, eligibility cut-offs, etc.)

3. Adjustment for Observable Differences

Four Solutions, continued

Variants on this approach include:

- Matching, Case-Control
- Regression
- Fixed effects (sibling/person as own control)
- propensity score

4. Instrumental Variable

Suppose you find and instrument (Z) that is:

- correlated with treatment: $E(Z'T) \neq 0$
- Uncorrelated with outcome, conditional on treatment: $E(Z'\varepsilon)=0$

Three examples of method (from my recent research projects)

- Did generous EITC benefits slow down gentrification in DC? We merged almost perfect administrative data and public data (Otabor, Kurban & Schmutz, 2019)
- MPDU owner program. We merged MPDU purchaser program data with public data. Through lottery units randomly allocated but it was not a perfect lottery system (Diagne, Kurban & Schmutz, 2018)
- MPDU rental program: Merged program data with public data. Units are not randomly allocated (Baglan, Kurban & McLeod, 2019)
- Synthetic micro samples at smaller geography (from PUMA to census tracts).
- Randomly allocated PUMA level observations to all census tracts and created census tract level micro samples by using census tract level distributions of 52 variables (Kurban et al 2011)

The Role of EITC on Migration within the District of Columbia

- Otabor, Kurban and Schmutz, (2019) Administrative Data

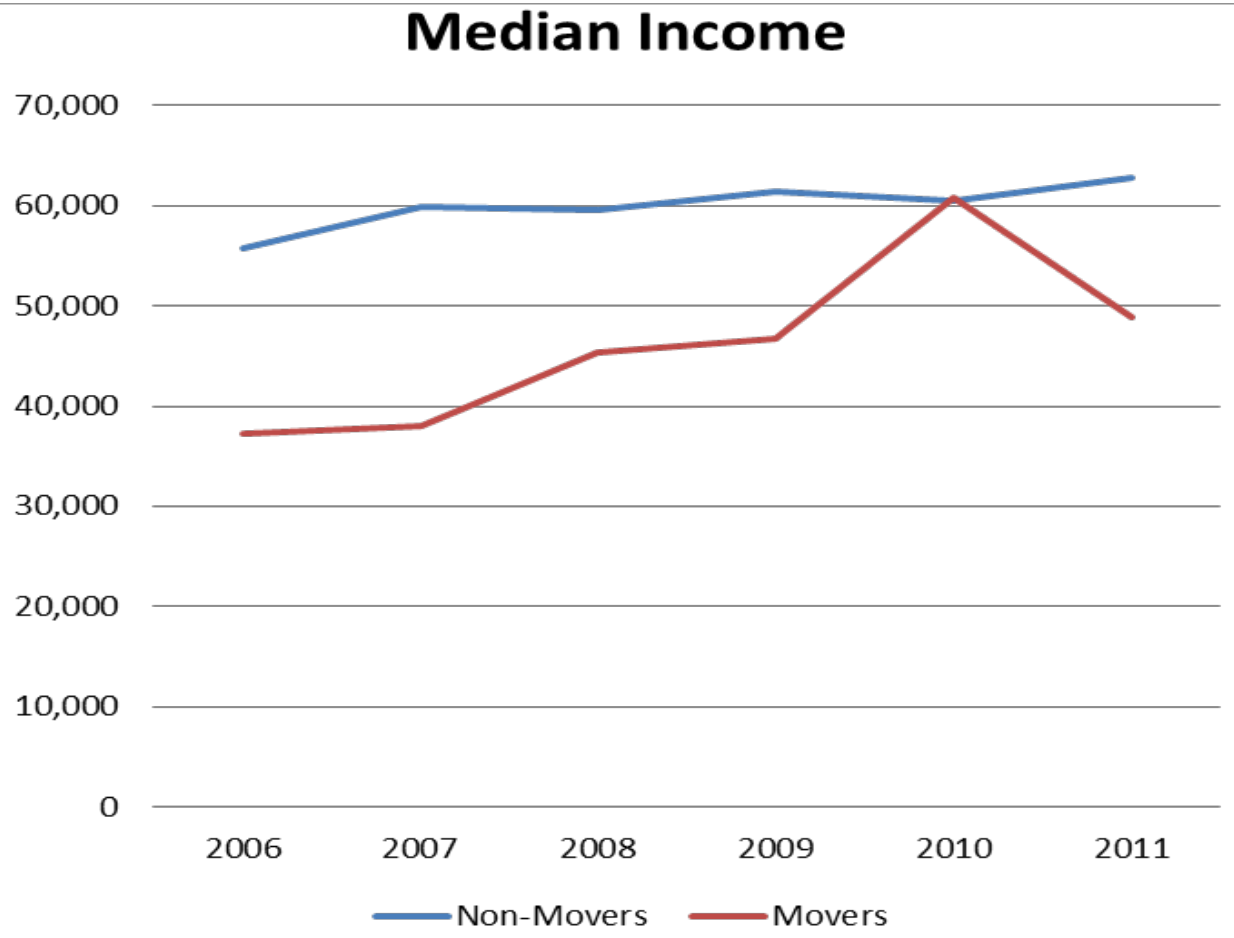
Methodology

- Poisson Pseudo-maximum-likelihood Estimator (PPML)
- Santos Silva, J.M.C. And Tenreyro, Silvana (2006); Chort And Rupelle (2015)

Data: 2005-2011

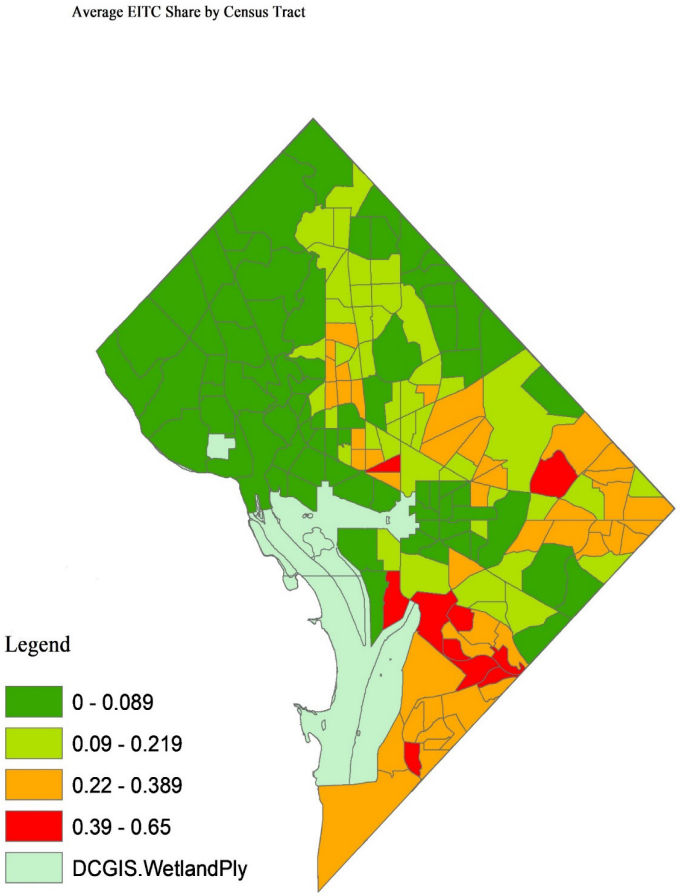
- Individual Income Tax And Real Property Tax Data (2005-2006, 2006-2007, 2007-2008, 2008-2009, 2009-2010, 2010-2011)
- American Community Survey (ACS)
- Neighborhoodinfodc

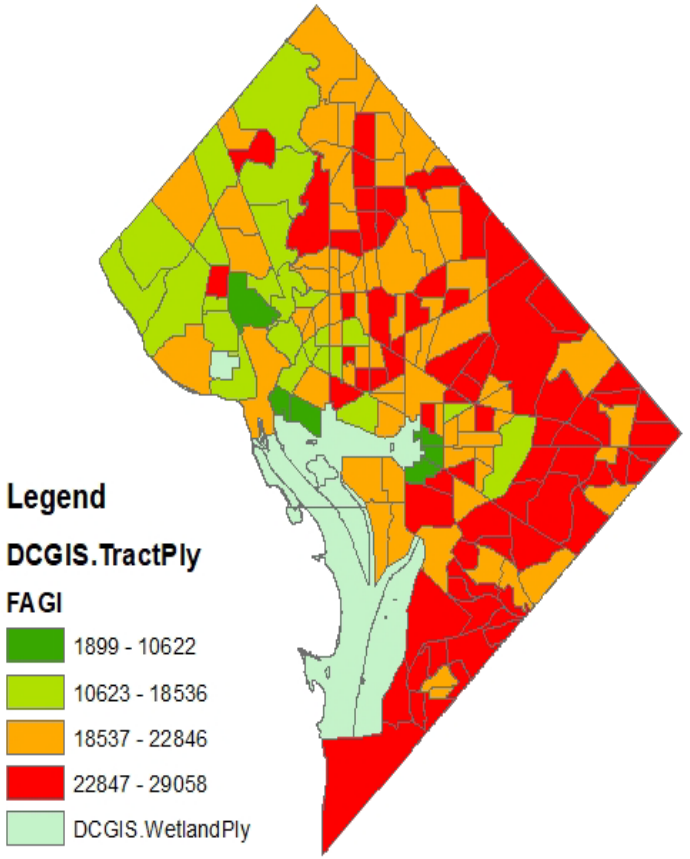
Median Income of all Movers within Washington, D.C.



Source: D.C 2005-2011 individual

Demographics- EITC Recipients by Census Tract 2005-2011





Did MPDU owner program benefit all?

- Diagne, Kurban & Schmutz (2018)

  (1) Does the MPDU purchaser program equitably allocate housing units among its applicants?

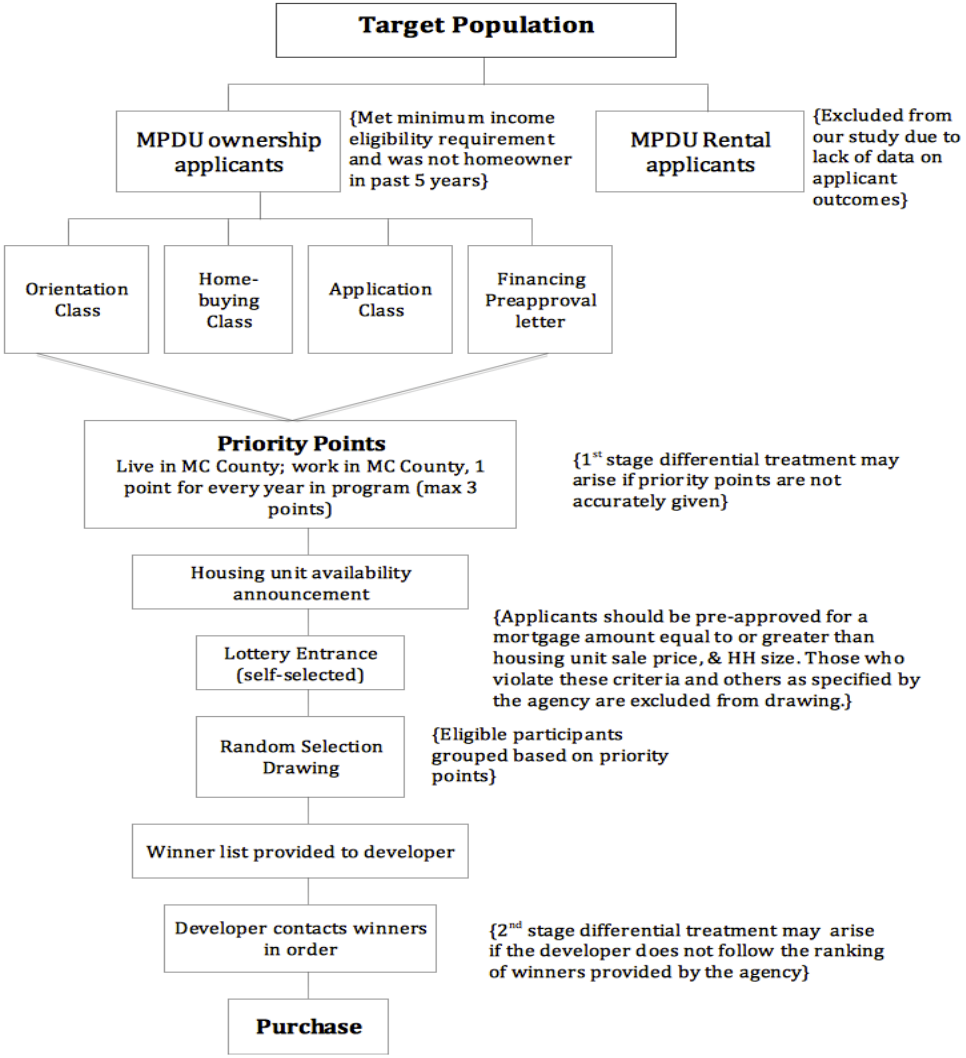
  (2) Is the program implemented as designed?

- Appropriate Methods:

a) Propensity Score Matching

b) Hedonic and logistic regressions

c) Sorting Indices to measure racial integration


Target Population
MPDU ownership applicants
{Met minimum income eligibility requirement and was not homeowner in past 5 years}
MPDU Rental applicants
{Excluded from our study due to lack of data on applicant outcomes}
Orientation Class
Home-buying Class
Application Class
Financing Preapproval letter
Priority Points
Live in MC County; work in MC County, 1 point for every year in program (max 3 points)
{1st stage differential treatment may arise if priority points are not accurately given}
Housing unit availability announcement
Lottery Entrance (self-selected)
{Applicants should be pre-approved for a mortgage amount equal to or greater than housing unit sale price, & HH size. Those who violate these criteria and others as specified by the agency are excluded from drawing.}
Random Selection Drawing
{Eligible participants grouped based on priority points}
Winner list provided to developer
Developer contacts winners in order
{2nd stage differential treatment may arise if the developer does not follow the ranking of winners provided by the agency}
Purchase

Did the MPDU Rental Housing Program in MC Improve Access to Affordable Housing?

Baglan, McLeod &Kurban, (2019)

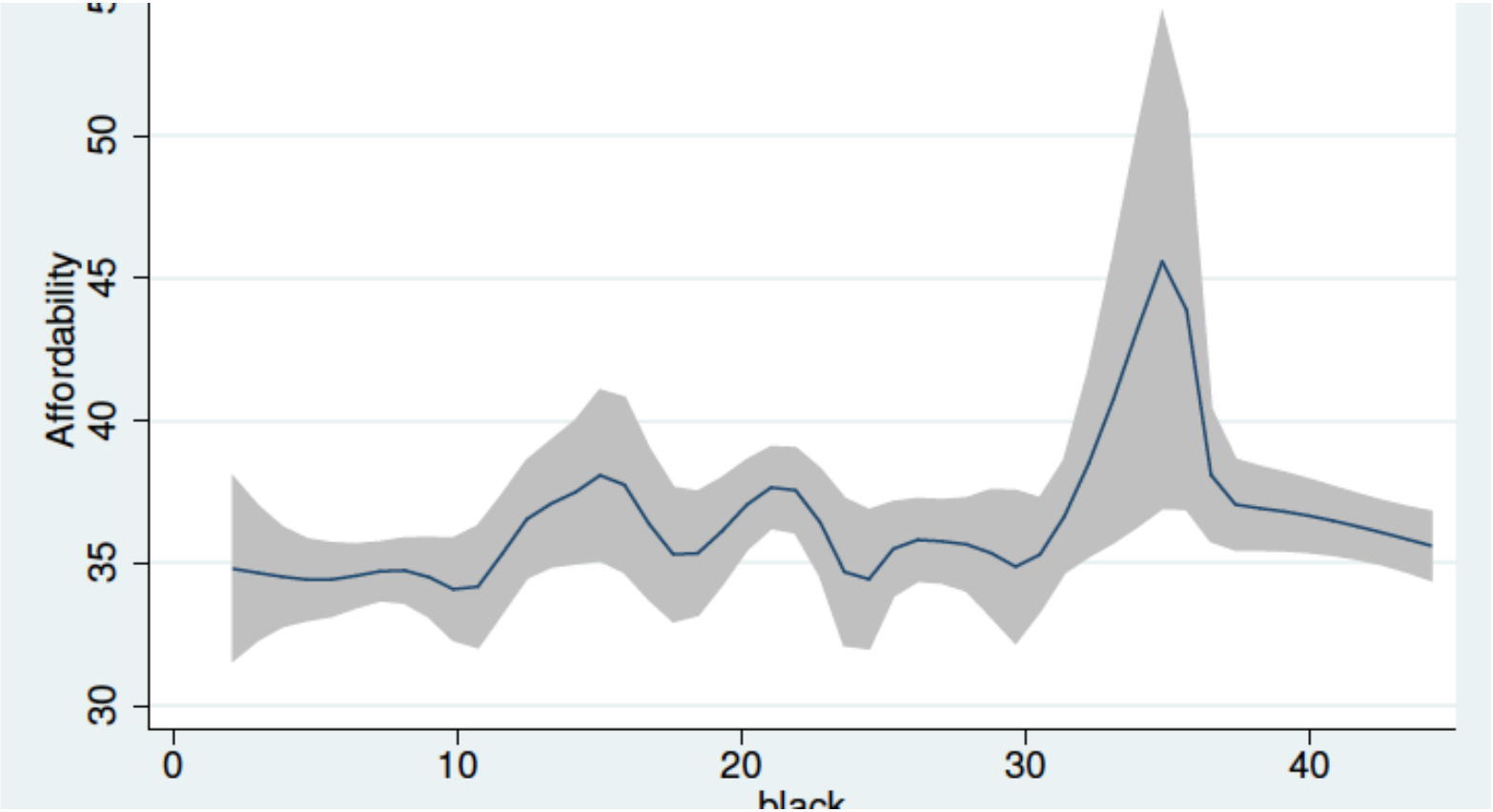
- **Data**: Rental contracts. 750 observations.
- 2008-2018. 74% of the observations between 2014 and 2018.
- Rental contract have the address, household size, household income, number of bedrooms, rental rate.
- Income limits and rent limits are provided by the Montgomery County.
- **Merge with neighborhood level vars**: Black pop. share, Hispanic pop. share, Median Household Income, Elementary School Ranking, Unemployment rate, Poverty rate.
- Limited Data: Race or immigrant status of the beneficiaries not known

Linear Regression as Appropriate Method

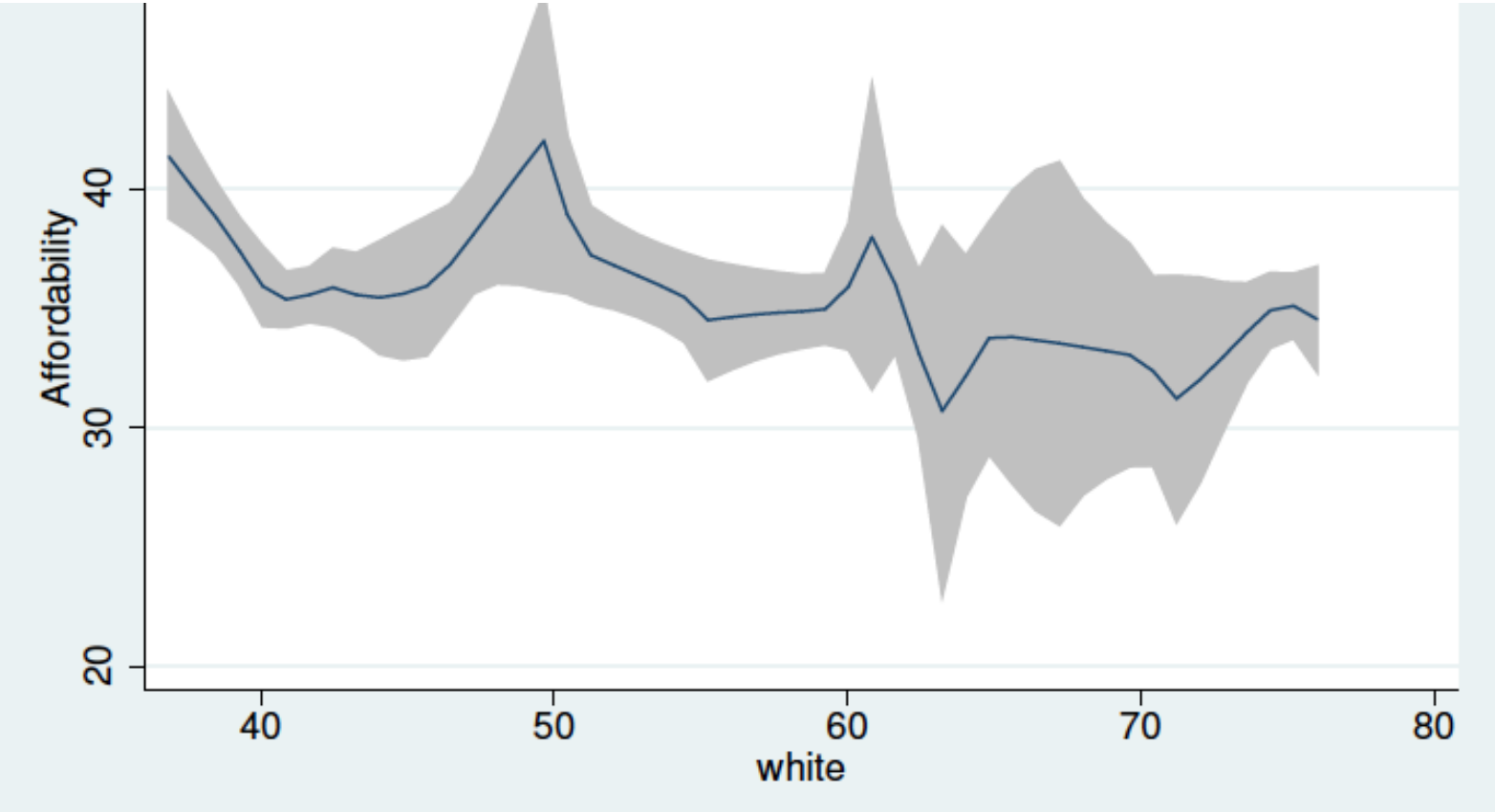
$$Affordability_{ij} = \beta_0 + \gamma' x_i + \delta' z_j + u_{ij},$$

- x_i: Household size, Number of bedrooms, Rental limit.
- z_j: control variables for urban characteristics at the census tract level.
  - Percentage of Black
  - Percentage of Hispanic
  - Unemployment rate
  - Median Household Income
  - Elementary School Ranking
  - Pecentage of Households Below Poverty
  - Educational Attainment

Affordability Index 100*Rent/Income



Affordability Index, White %



Incomplete Data

- Our data suggests that program participants are lower income but
- Relatively higher income participants choose white neighborhoods
- MPDU Rental houses appear to be less affordable in white neighborhoods
- Why?
- Higher income participants are willing to pay higher share of income to have access to better neighborhoods
- We do not know race of program participants
- How to get around this problem?
- Any suggestion?

Other Data Issues

- Limitations of public data

- Privacy issues limit data availability at finer geography (example: Census tract or block groups)

- Privacy issues limit availability of some variables (top coding, grouping and missing observations)

- Remedies: Use GIS to combine data from different geographic details (example: concentration of crime incidences, fast food places around neighborhoods

-) Creating and using Synthetic Data Sets (example, we created Census block group level Micro Samples by using heuristic methods such as hill climbing and proportional fitting procedure in Kurban et al 2012.

Exhibit 3a

PUMA 3101 Real Cross-Tabulations

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	111	23	8	2	3
50,000–79,999	132	28	6	7	3
80,000–89,999	82	29	19	4	3
90,000–99,999	115	18	21	5	2
100,000–124,999	223	60	47	15	6
125,000–149,999	247	55	35	15	4
150,000–174,999	146	44	39	11	8
175,000–199,999	113	34	24	10	1
200,000–249,999	90	26	30	12	6
250,000–299,999	60	16	13	5	5
300,000–399,999	37	9	11	0	1
400,000–499,999	9	0	3	1	0
500,000+	16	2	3	1	0

PUMA = Public Use Microdata Area.

Exhibit 3c

IPF Cross-Tabulations for PUMA 3101*

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	106	24	10	4	3
50,000–79,999	124	28	12	8	4
80,000–89,999	94	21	17	4	2
90,000–99,999	115	22	17	5	2
100,000–124,999	234	58	42	11	5
125,000–149,999	238	55	41	14	8
150,000–174,999	158	40	33	11	5
175,000–199,999	113	32	26	9	3
200,000–249,999	93	29	29	9	5
250,000–299,999	58	18	14	6	3
300,000–399,999	30	11	10	5	2
400,000–499,999	7	3	2	1	1
500,000+	12	3	4	2	0

IPF = iterated proportional fitting. PUMA = Public Use Microdata Area.

** Values are rounded to the nearest whole number.*

Exhibit 3b

Hill-Climbing Cross-Tabulations for PUMA 3101*

House Value ($)	Number of Children				
	0	1	2	3	4+
0–49,999	110	24	7	2	4
50,000–79,999	128	29	8	8	3
80,000–89,999	94	20	21	2	0
90,000–99,999	116	21	15	8	1
100,000–124,999	236	63	35	12	5
125,000–149,999	243	48	43	14	8
150,000–174,999	153	39	40	11	5
175,000–199,999	109	35	30	5	3
200,000–249,999	98	24	25	11	6
250,000–299,999	55	21	13	6	4
300,000–399,999	24	13	12	6	3
400,000–499,999	3	4	4	2	0
500,000+	12	3	6	1	0

PUMA = Public Use Microdata Area.

** Values are rounded to the nearest whole number.*

Going beyond public use data and administrative data

- Data scraping (extracting data from websites)
- Many research papers and new dissertations scrap data from various websites (example: what type of restaurants survive in cities? Scrap menu and demand from restaurant websites)
- Scrap data from google search, facebook and twitter (example: assessing public sentiments during an event such as natural disaster, elections, or big demonstrations)
- Big Data tools: R and beyond (Example: We extracted 3-day and 7-day local weather forecast data from National Weather Service by using R)
- Increasingly Census Bureau and other data sets are supplemented by R codes. One can create variables and perform analysis by using a comprehensive R script.

Big Data and Poverty

Constructing spatiotemporal poverty indices from big data☆

Christopher Njuguna [a,*], Patrick McSharry [a,b,c]

[a] ICT Center of Excellence, Carnegie Mellon University, Kigali, Rwanda
[b] Smith School of Enterprise and the Environment, University of Oxford, UK
[c] Oxford Man Institute of Quantitative Finance, University of Oxford, UK

ARTICLE INFO

Available online 16 August 2016

Keywords:
Call detail record (CDR)
Poverty index
Machine learning
Big data
Socioeconomic level
Rwanda

ABSTRACT

Big data offers the potential to calculate timely estimates of the socioeconomic development of a region. Mobile telephone activity provides an enormous wealth of information that can be utilized alongside household surveys. Estimates of poverty and wealth rely on the calculation of features from call detail records (CDRs), however, mobile network operators are reluctant to provide access to CDRs due to commercial and privacy concerns. As a compromise, this study shows that a sparse CDR dataset combined with other publicly available datasets based on satellite imagery can yield competitive results. In particular, a model is built using two CDR-based features, mobile ownership per capita and call volume per phone, combined with normalized satellite nightlight data and population density, to estimate the multi-dimensional poverty index (MPI) at the sector level in Rwanda. This model accurately estimates the MPI for sectors in Rwanda that contain mobile phone cell towers (cross-validated correlation of 0.88).

Public Data Sources

- U.S. Census (https://www.census.gov/)
  - CPS (https://www.census.gov/cps/data/), various supplements
  - ACS (https://www.census.gov/programs-surveys/acs/)
  - SIPP (https://www.census.gov/sipp/)
  - BLS (https://www.bls.gov/)
  - HUD (https://www.huduser.gov)
  - IPUMS.org
- Board of Governors of Federal Reserve System (www.federalreserve.gov)
  - Survey of Consumer Finances (SCF)
  - Survey of Household Economics and Decision-making(SHED)

Longitudinal

- Panel Study of Income Dynamics (https://psidonline.isr.umich.edu/)
- Fragile Families (https://fragilefamilies.princeton.edu/)
- National Longitudinal Survey of Youth (https://www.bls.gov/nls/nlsy79.htm)
- National Longitudinal Study of Adolescent to Adult Health (Add Health) (http://www.cpc.unc.edu/projects/addhealth)
- Early Childhood Longitudinal Survey, Birth Cohort (ECLS-B) (https://nces.ed.gov/ecls/birth.asp)
- **Administrative Data**
- Federal, state, local, private sector (county, cities, villages, companies) collect data
- Example: Moderately Priced Dwelling Units (MPDU), Montgomery County
- DC government tax data (income and property tax data)

References

- A Beginner's Guide to Creating Small Area Cross Tabulations, H Kurban, R Gallagher, GA Kurban, J Persky - Cityscape, 2011.
- Demographics of Payday Lending in Oklahoma, Haydar Kurban and Adji Diagne 2014.
- http://coas.howard.edu/centeronraceandwealth/reports&publications/Oklahoma%20Payday%20Lending%20Report%20Final%20For%20Website.pdf
- Ybara, Marci, Quantitative Analysis, Summer Dissertation Workshop Proposal, 2018, Howard University