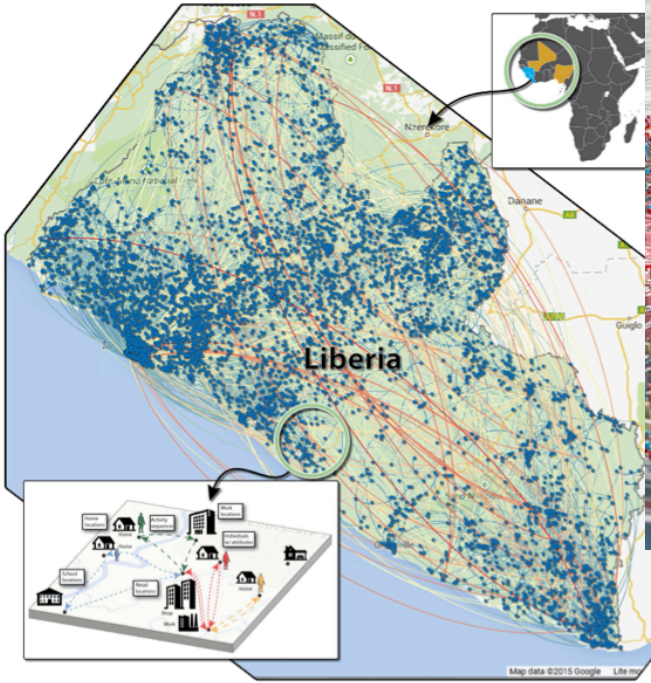
Combining federal, administrative, and local data for community surveillance

Josh Goldstein, SDAL, Virginia Tech
Emily Molfino, US Census
Dave Higdon SDAL, Virginia Tech
Sallie Keller, SDAL, Virginia Tech
Stephanie Shipp, SDAL, Virginia Tech
Shawn Buckholtz, US HUD

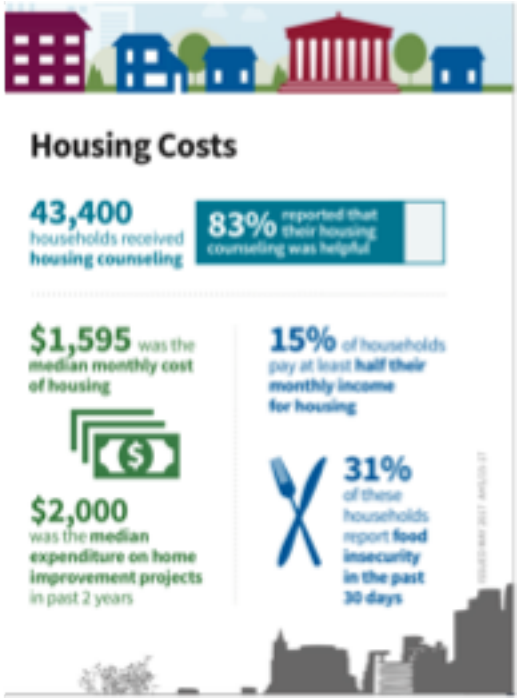
Synthetic Populations

Simulations and Agent-based Models



Planning

Assessing Surveys



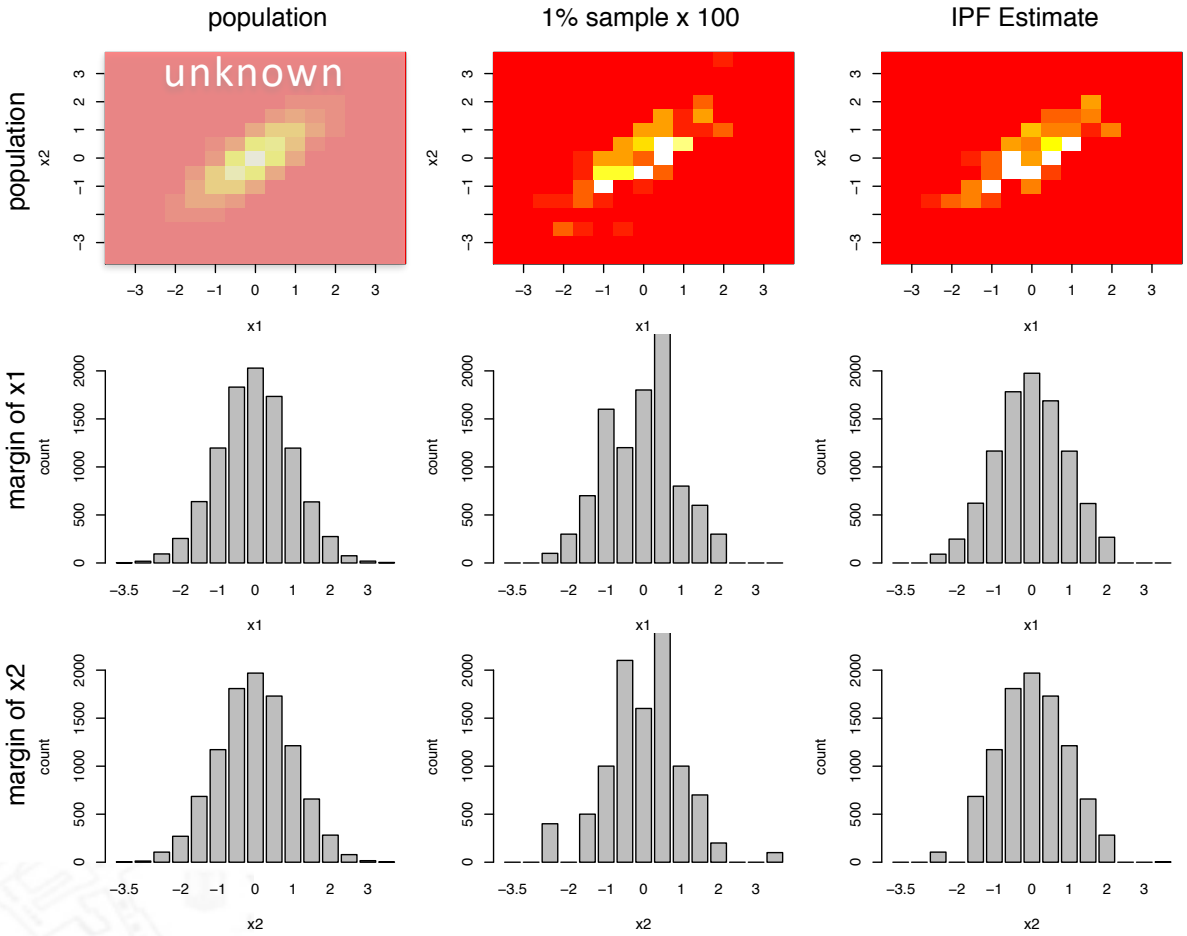
Standard Approaches for Constructing a Synthetic Population

Approaches

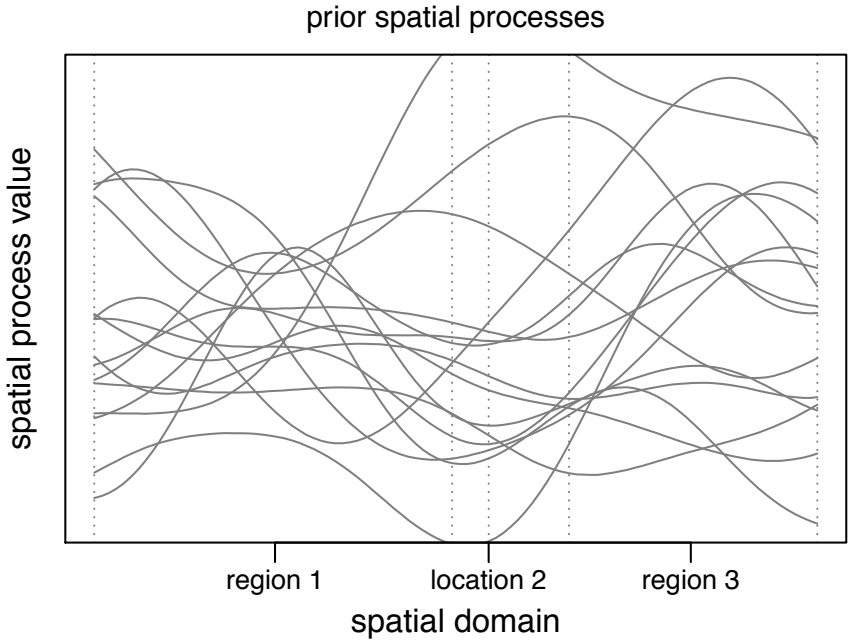
- Iterative proportional fitting (IPF)
- Annealing
- Model-based

Issues

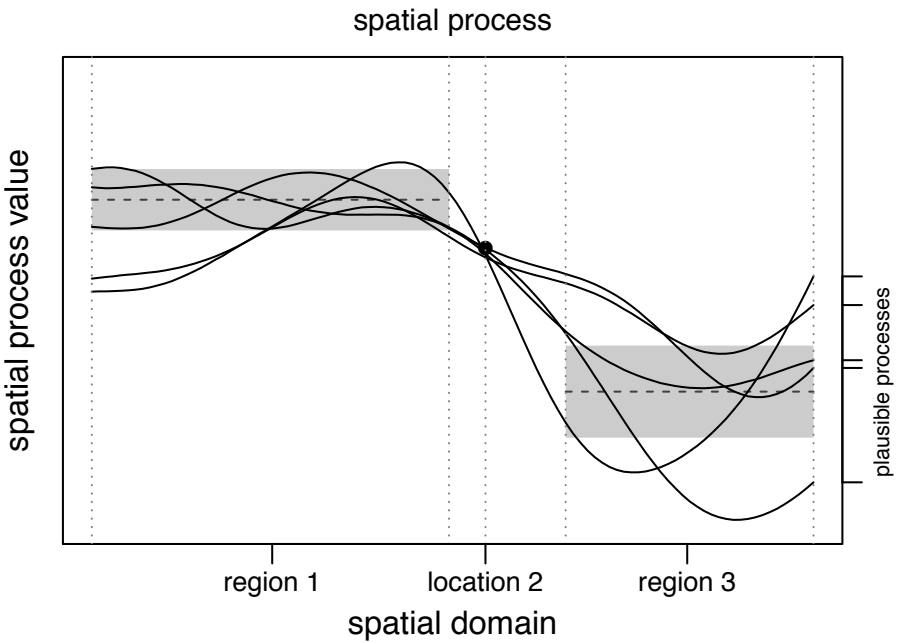
- Often replicates individuals
- Solution not unique
- Uncertainty not well characterized



Spatial Models:
Uncertainty & Data at Different Levels of Aggregation

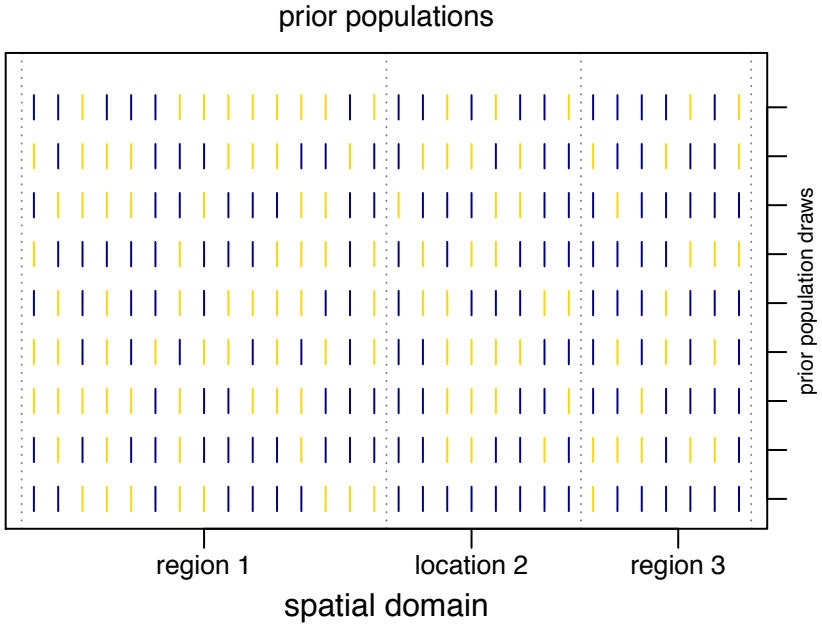


The prior model for the spatial process specifies the spatial dependence and variability that is appropriate for this process.

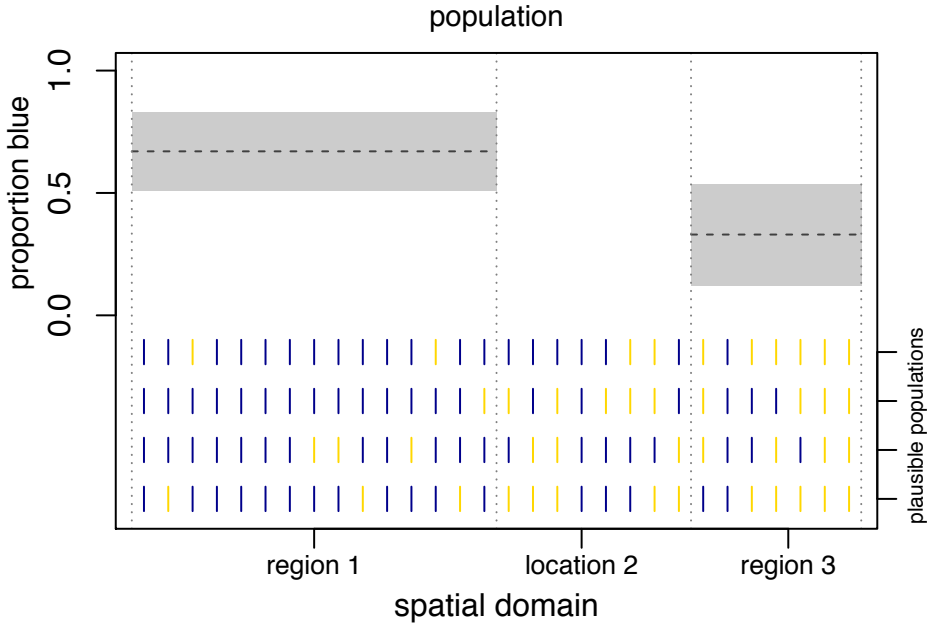


Spatially linked data provide additional information about the spatial process: the average over spatial region 1; the value at location 2; and the average over spatial region 3.

Population Models: Data at Different Levels of Aggregation

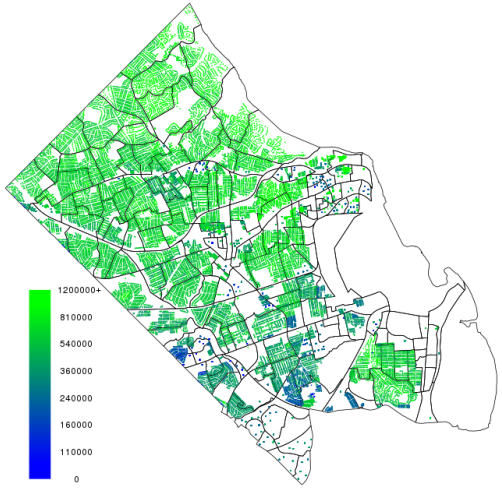


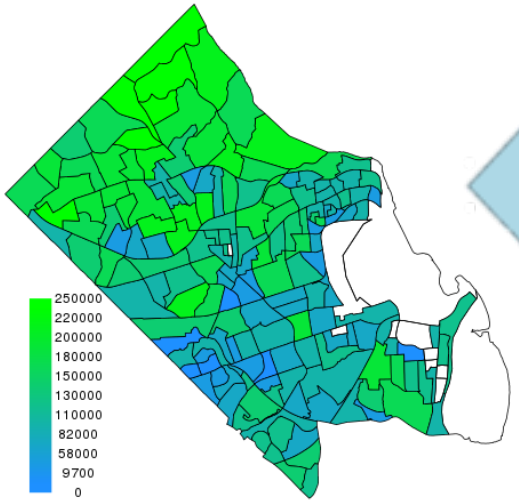
The prior model specifies the spatial location of each individual and requires the overall proportion of blue individuals to be near .6 to be consistent with survey data from this population.

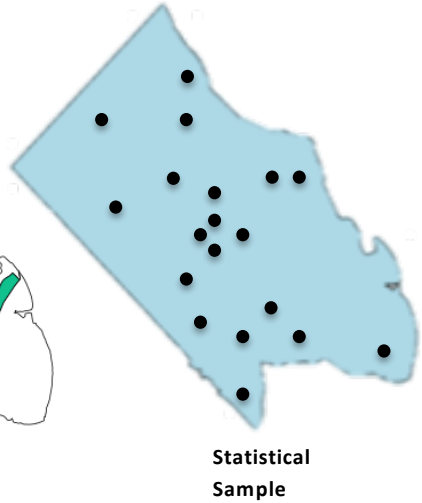


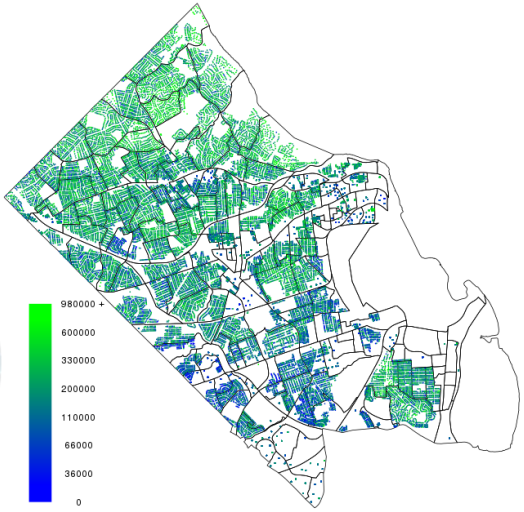
Spatially linked data provide additional information about the proportion: proportion of blue individuals in spatial region 1; the color of the individual at location 2; and the proportion of blue individuals in spatial region 3.

Use data at different levels of aggregation / support to inform at the individual level





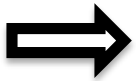




Local Data
housing price
tax amount from
real estate tax
assessments

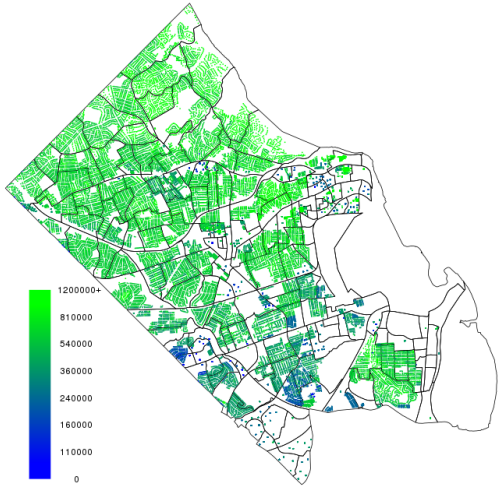
ACS
Income summary
- by block group

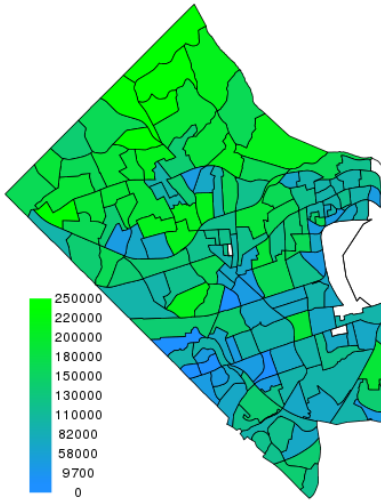
ACS Microdata
housing price
tax amount
income
- county sample

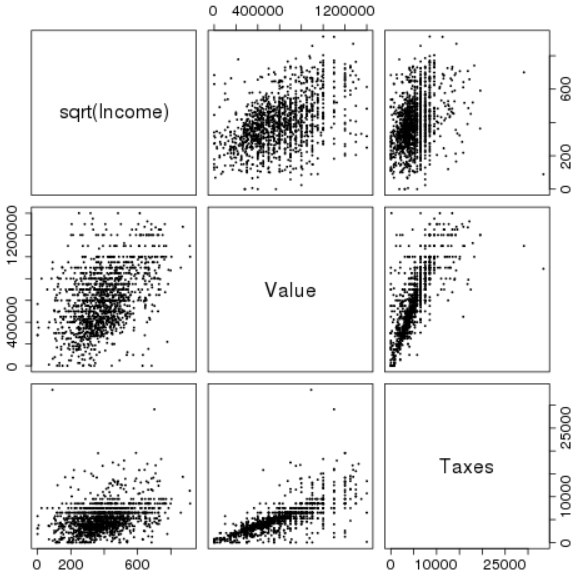


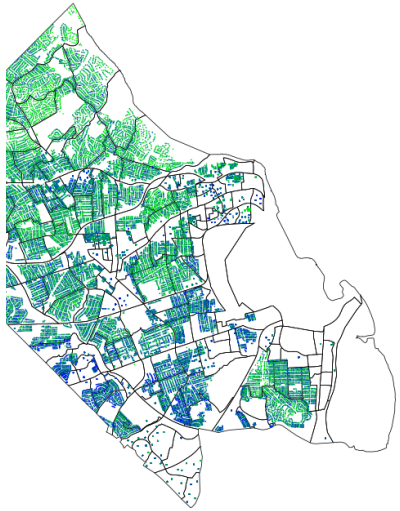
Household level Imputation
housing price
tax amount
income

Use data at different levels of aggregation / support to inform at the individual level









Local Data
housing price
tax amount from
real estate tax
assessments

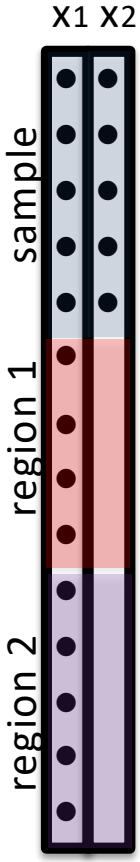
ACS
Income summary
- by block group

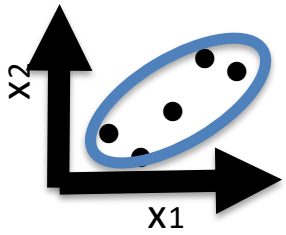
ACS Microdata
housing price
tax amount
income
- county sample

⇨

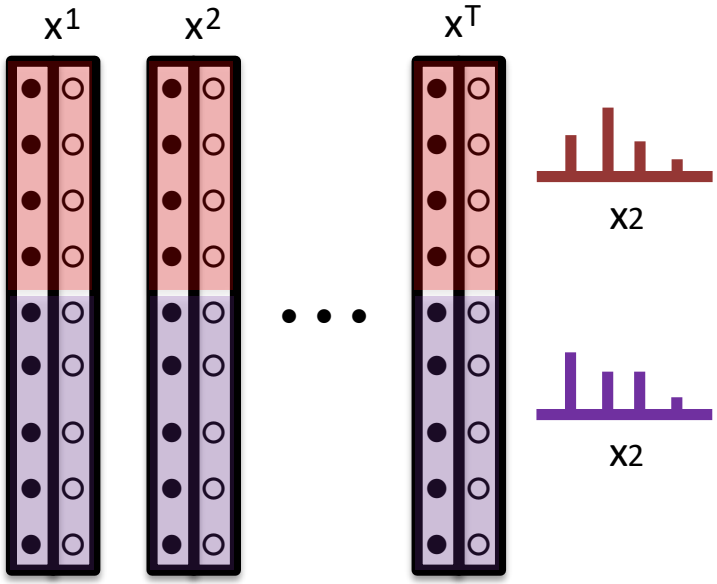
Household level Imputation
housing price
tax amount
income

“Posterior” and a Sampling Approach





$$\pi_{MI}(x_2|x_1) \times \pi(x_2|r)$$



draws from π_{MI}

Use M-H update scheme:

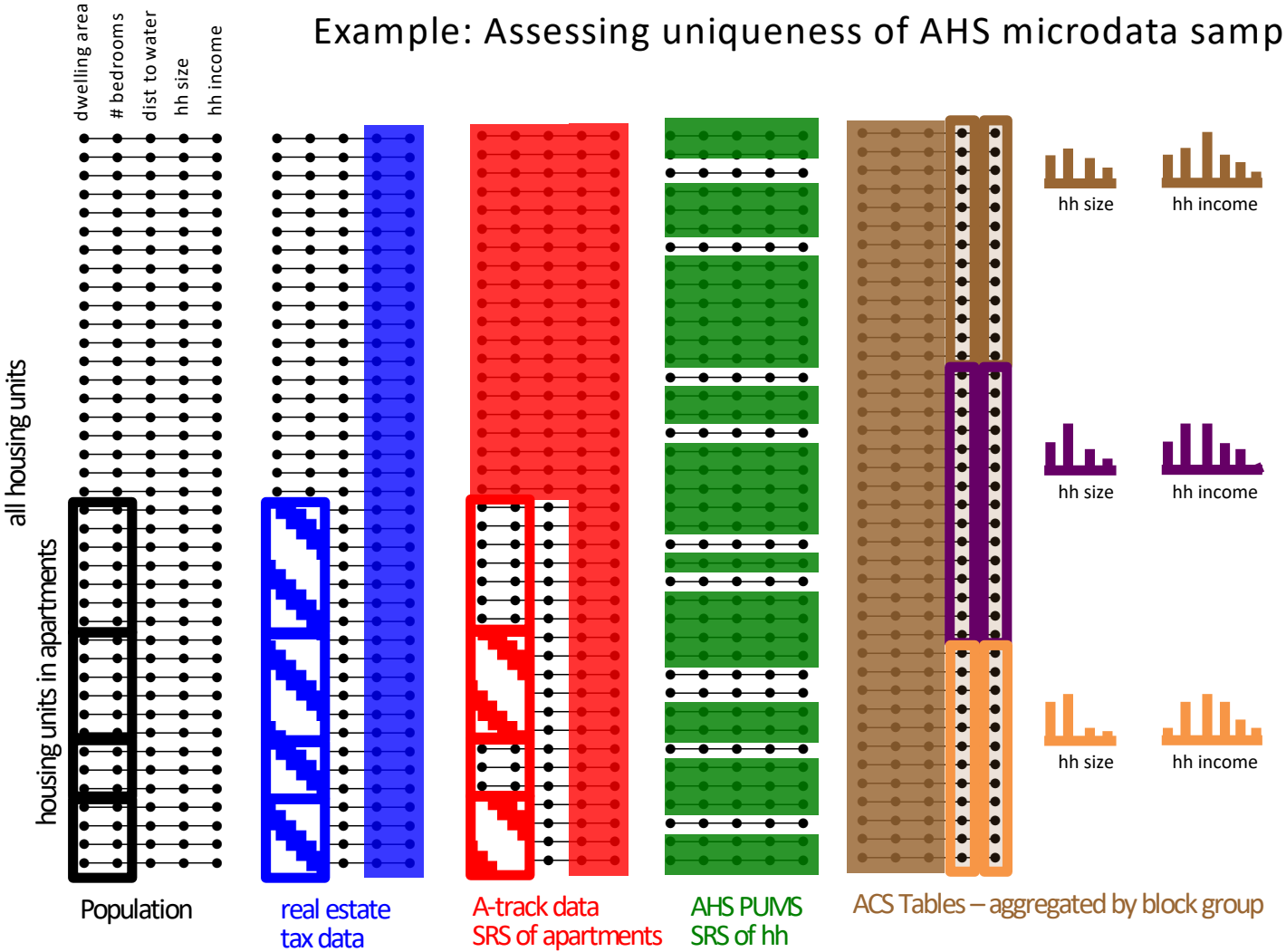
given x^t

draw $x'_{2k} \sim \pi_{MI}(x_2|x_1)$

$$\text{compute } \alpha = \frac{\pi_{MI}(x'_2|x_1)\pi(x'_2|r) \cdot \pi_{MI}(x_2|x_1)}{\pi_{MI}(x_2|x_1)\pi(x_2|r) \cdot \pi_{MI}(x'_2|x_1)}$$

$$\text{set } x_{2k}^{t+1} = \begin{cases} x'_{2k} \text{ with prob } \alpha \wedge 1 \\ x_{2k} \text{ with prob } 1 - \alpha \wedge 1 \end{cases}$$

Example: Assessing uniqueness of AHS microdata samples



Variables Used to Inform Population Distribution

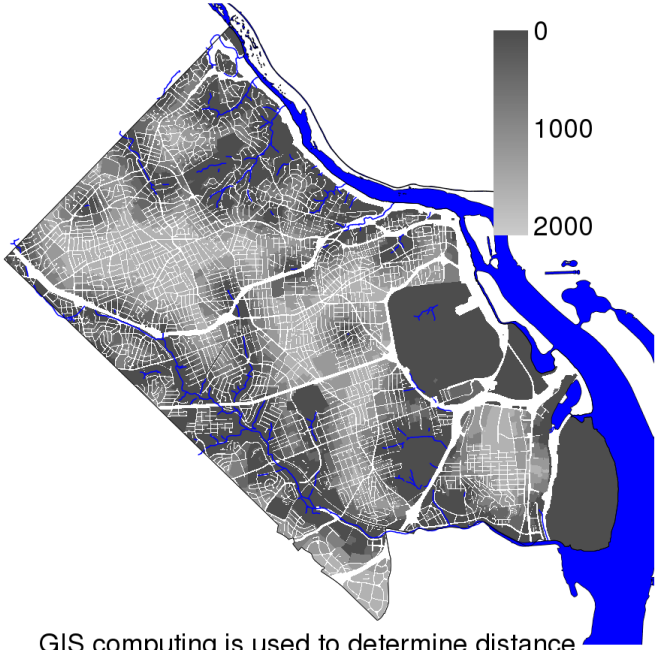
Renters

- Household Income
- Bedrooms
- Rent
- Year Built
- Within a half block of water

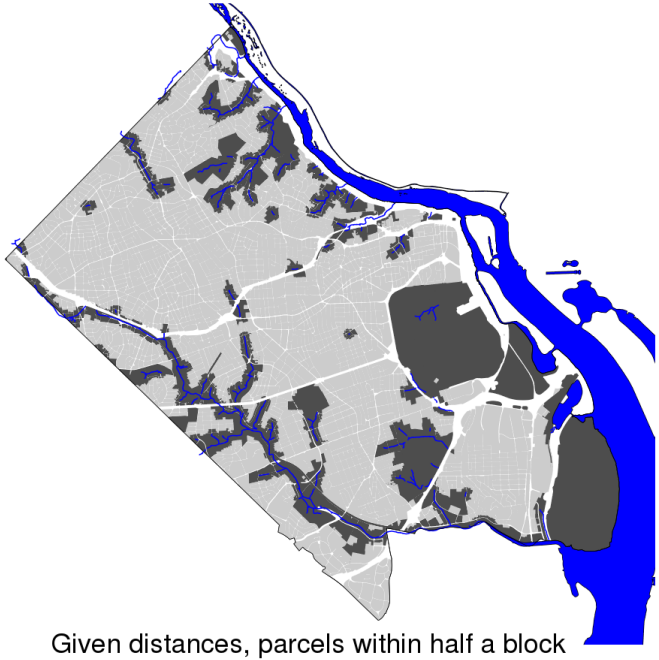
Owners

- Household Income
- Property Value
- Real Estate Taxes
- Bedrooms
- Bathrooms
- Living Area
- Year Built
- Within a half block of water

Direct Determination of Geographic Variables

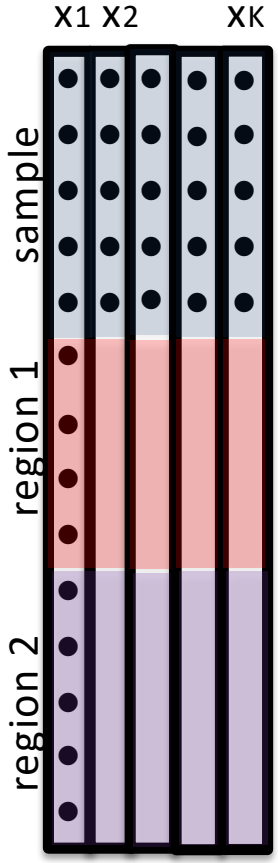


GIS computing is used to determine distance in feet to a body of water for each parcel in Arlington county.



Given distances, parcels within half a block of a body of water are shaded.

More Generally



tables

$\pi(x_j|r)$

Xj

Xj

apartments

$\pi(x_{j'}|s')$

opportunity data

$\pi(x_{j*}|s^*)$

•

•

•

survey data

¿ $\pi(x)$?

$\pi(x_1|x_{-1})$

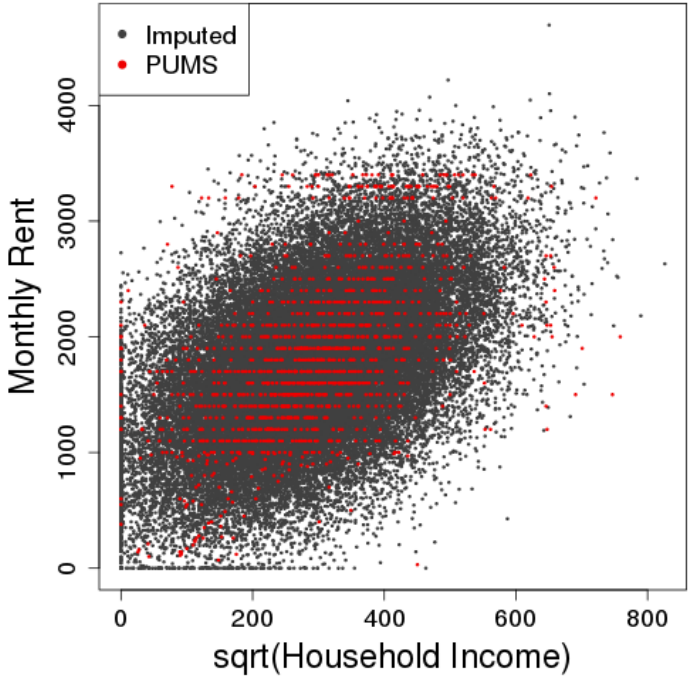
$\pi(x_2|x_{-2})$

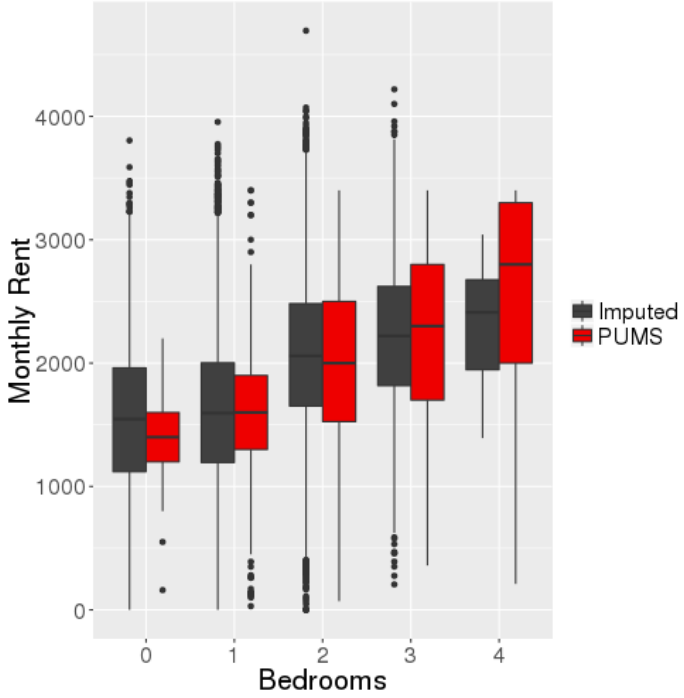
⋮

$\pi(x_K|x_{-K})$

Results – Conditional Distributions

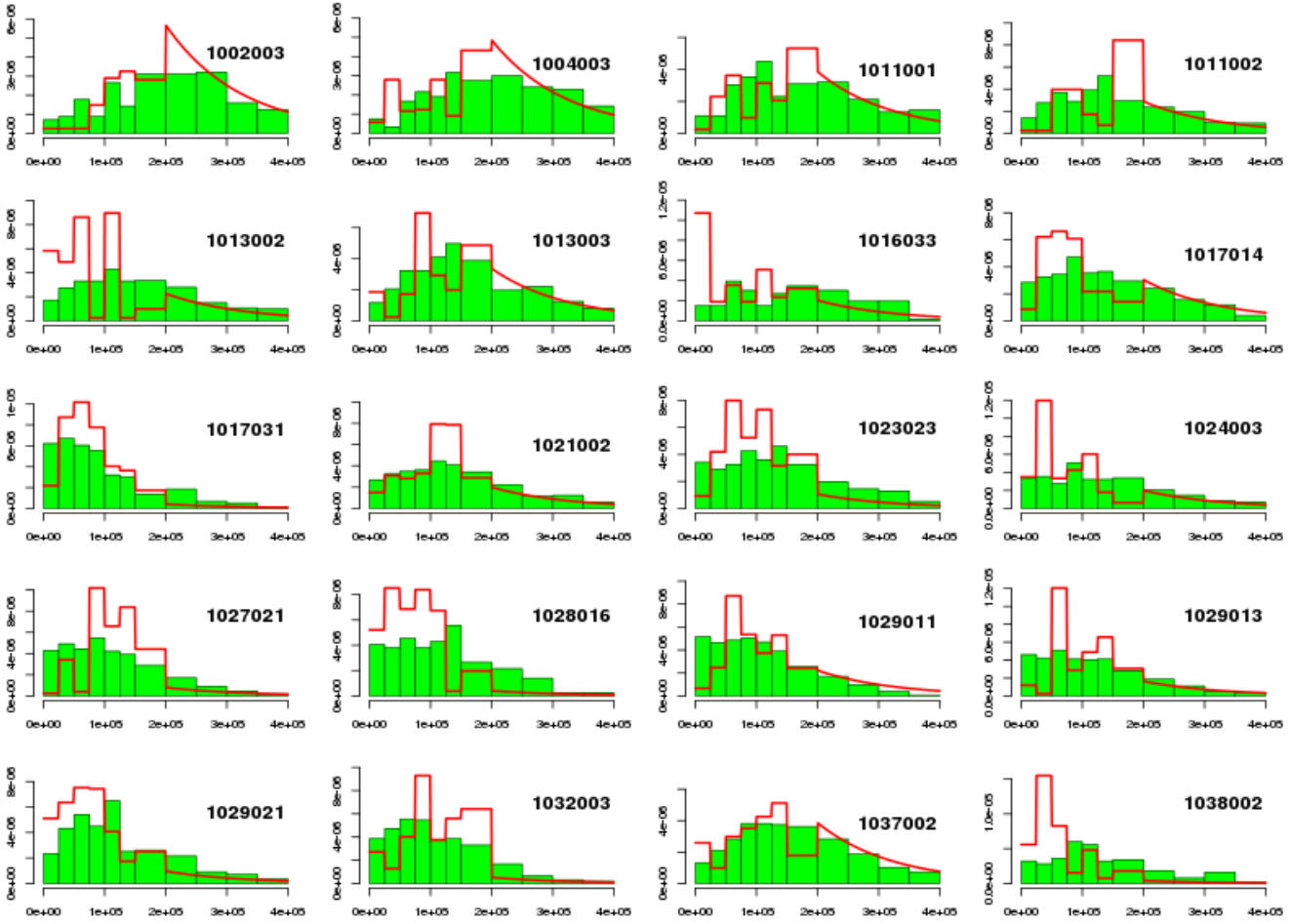
Imputed variables have conditional distributions consistent with those observed in the PUMS





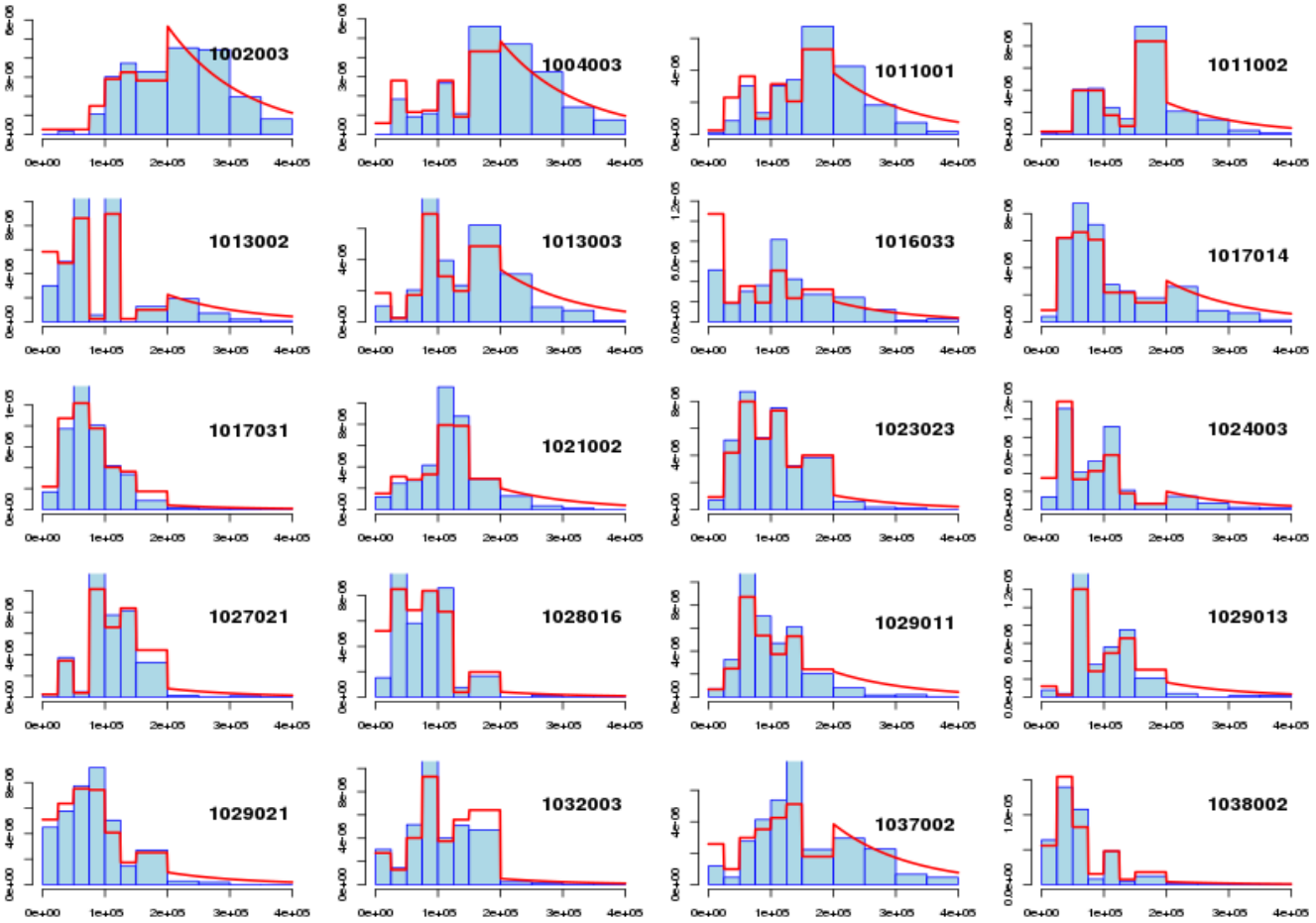
Using only microdata

household income

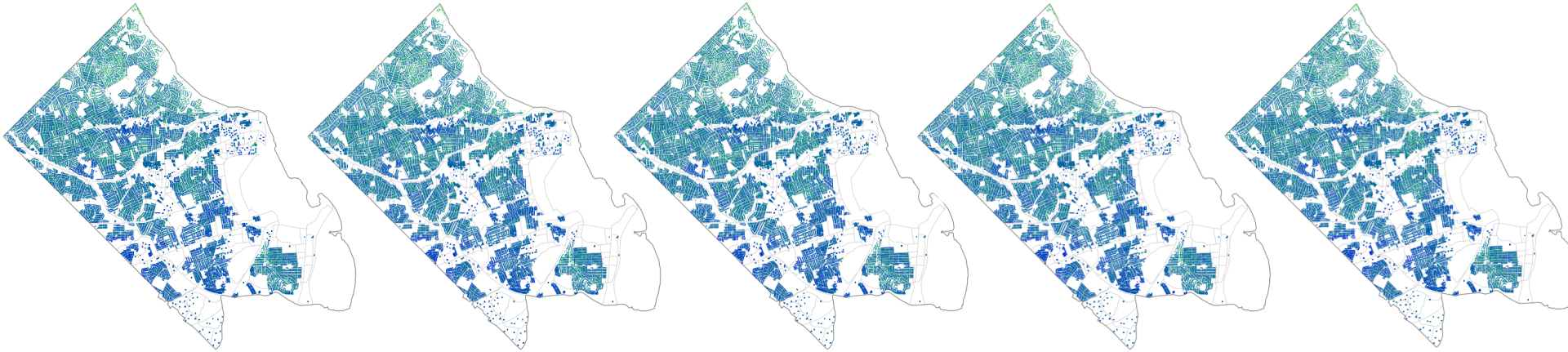


Also using marginal table information

household income



Posterior Realizations

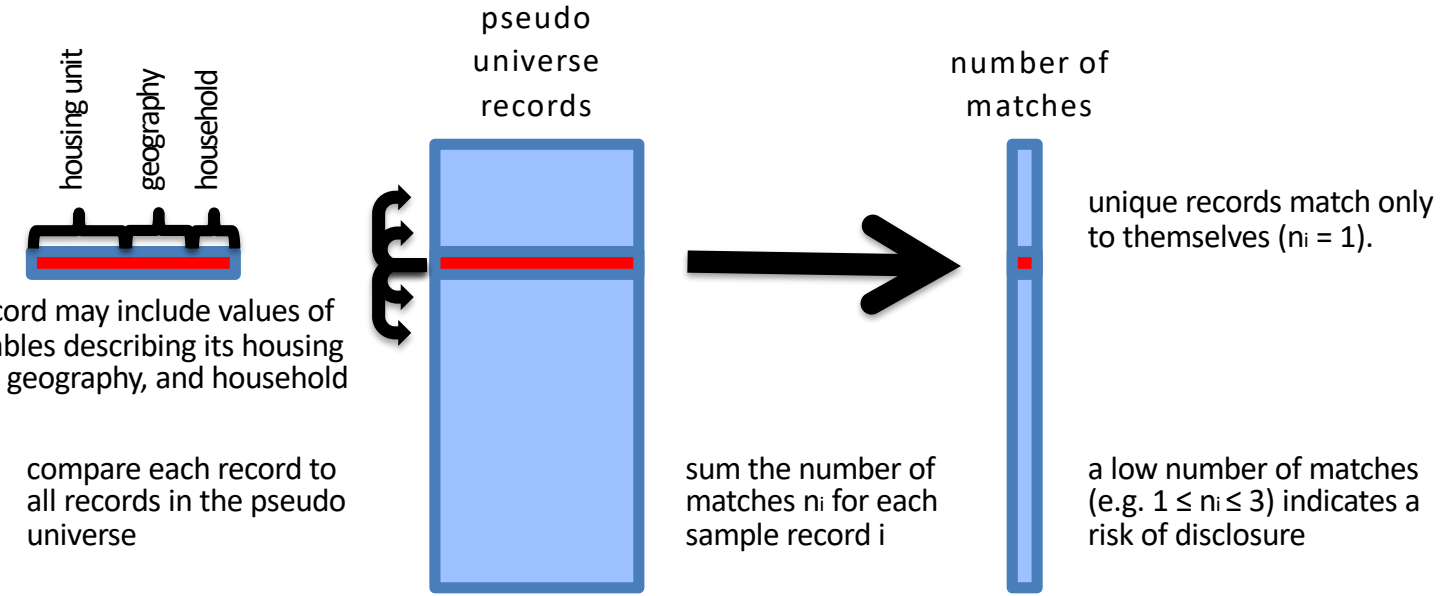


Realizations of household income

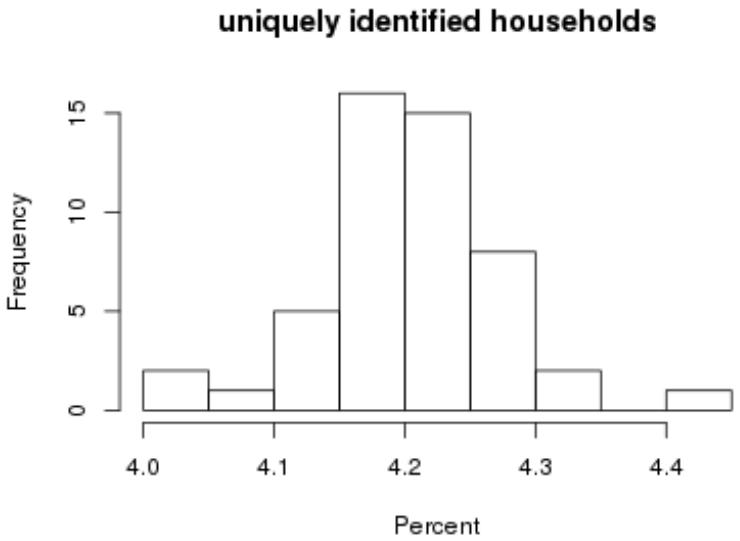
Uses information from:

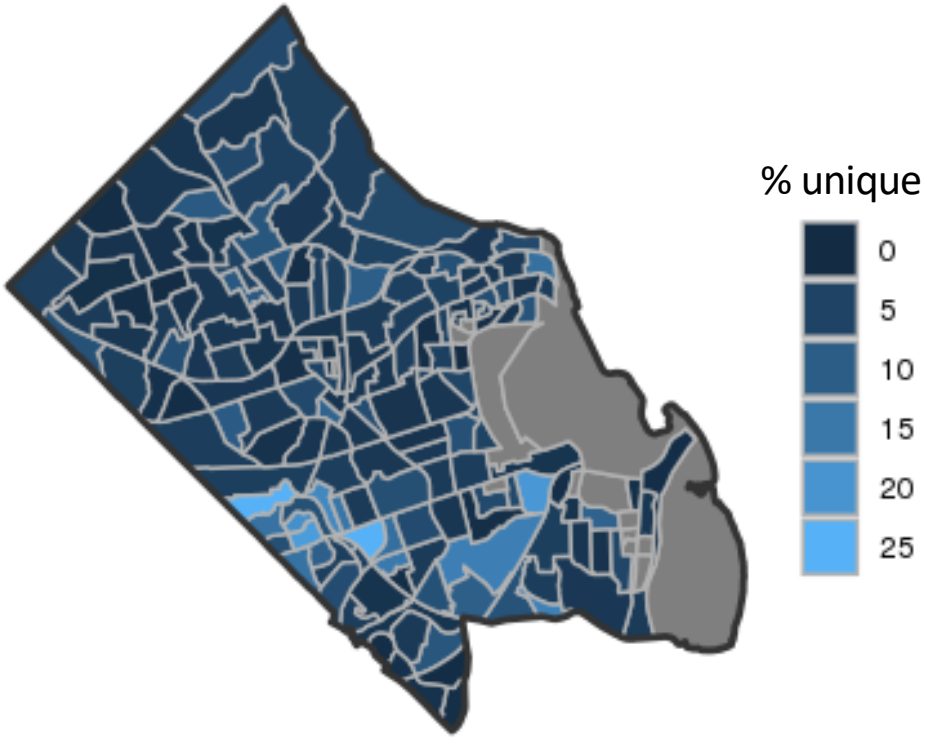
- real estate tax records,
- geography, county GIS data
- American Community Survey microdata,
- American Community Survey (blockgroup) summaries of:
  - household Income, property value, # bedrooms, year built
- County apartment surveys

Assessing uniqueness with these realizations of Arlington



Assessing uniqueness with these realizations of Arlington





Matching tolerances:

- Bedrooms, Bathrooms: 0-4 exact match, 5+ matches from 4-Inf
- Year Built: +/- 15 years
- Water Flag: Exact Match ("1" or "0")
- All the other continuous variables: +/- 25%

Our probabilities of identifiability are likely overstated

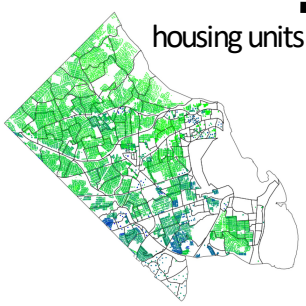
- respondents do tend to give answers to survey questions that do not match ours exactly
- Identifiability depends on tolerances

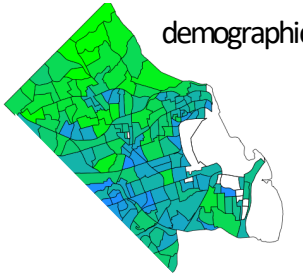
Discussion Points

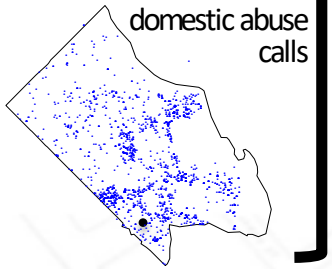
- Our probabilities of identifiability are likely overstated
  - respondents do tend to give answers to survey questions that do not match ours exactly
  - Identifiability depends on tolerances
- Pairing up population with point level events
  - e.g. local neighborhoods and domestic violence
- Change of support for aggregated data
  - e.g. combining aggregates by school district and political districts

Combining local data sources with a probabilistic representation of the local community

Multiple data sources at various levels of aggregation.







Combine using state of the art concepts from:

- Multiple Imputation
- Synthetic population generation
- Bayesian Computation

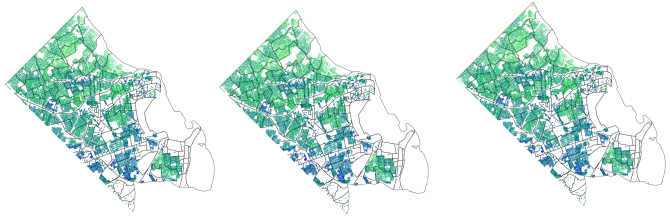
Generate plausible fine scale configurations of the population/system for prediction and quantifying uncertainty.

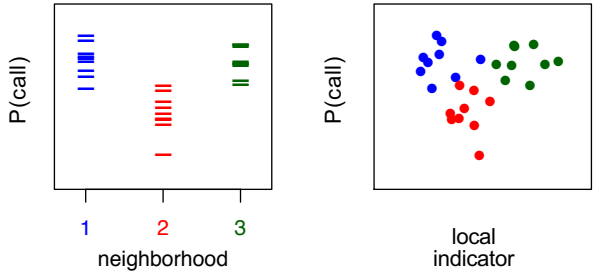
Local data sources relevant to understanding factors relating to domestic abuse calls:

- Real estate tax records
- Demographics (e.g. American Community Survey)
- Local Police Responses
- Local EMS Medical Responses

Posterior Realizations

- incorporate information and uncertainty in data combination

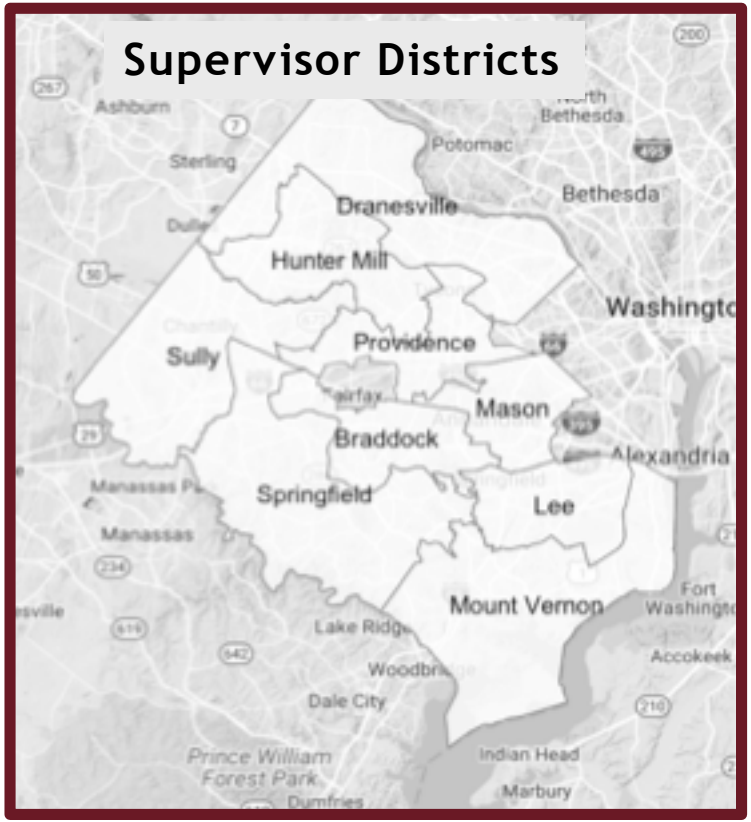




Statistically generated realizations allow separation of signal and noise at the local level.

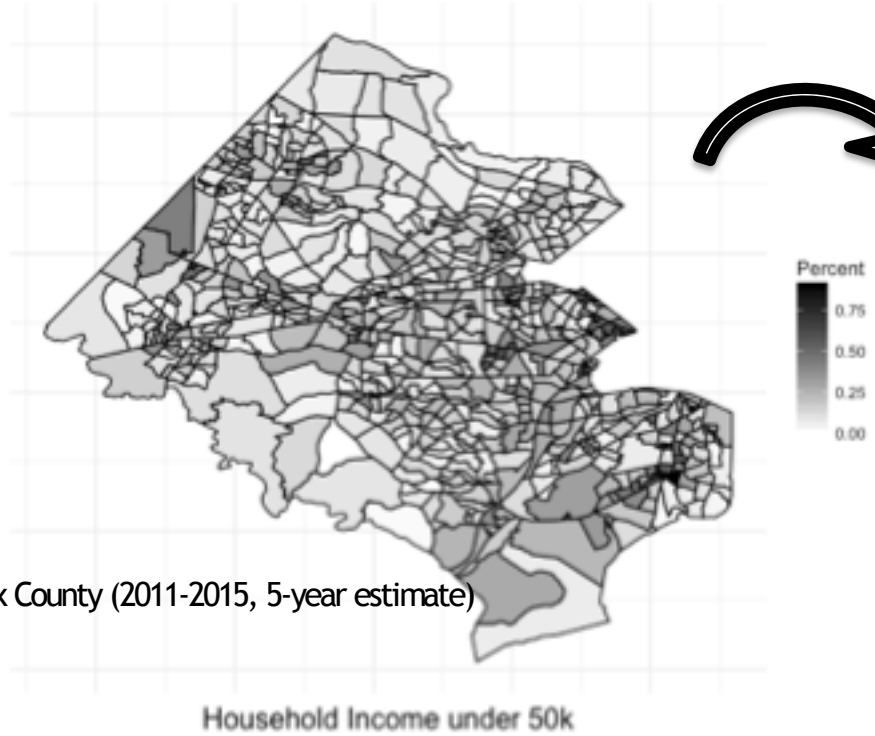
Example: Fairfax County, Virginia

Supervisor Districts and High School Attendance Areas



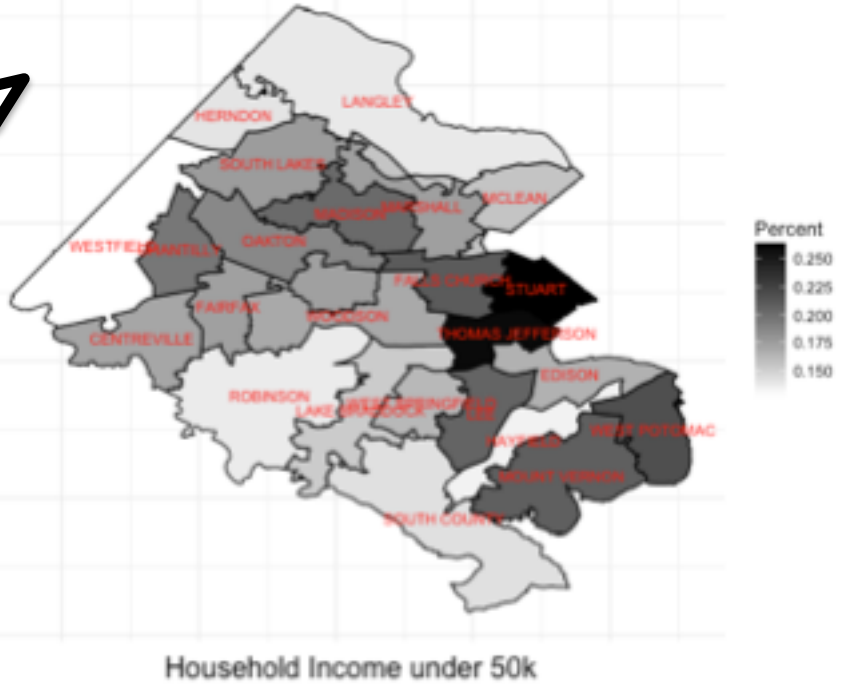
Change of support for aggregated data

Census Block Group



Household Income under 50k

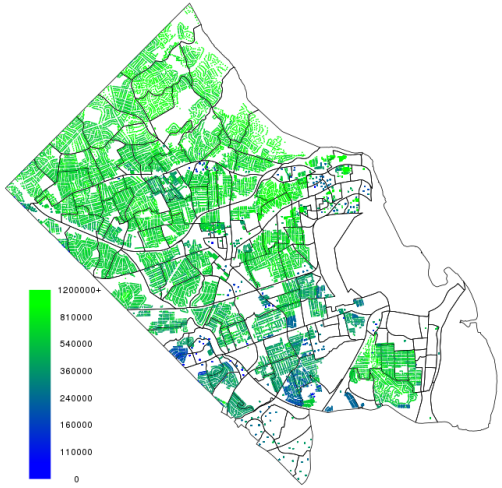
High School Pyramid Boundaries
(synthetic population)

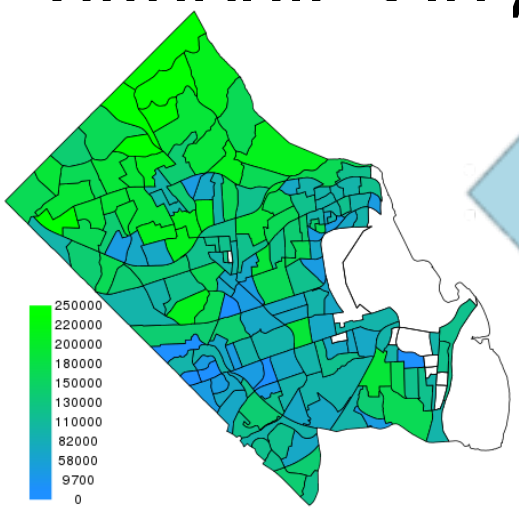


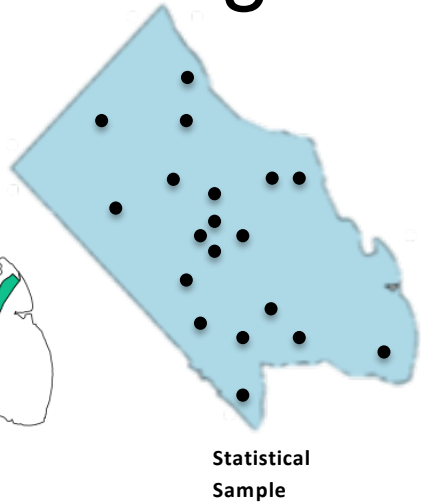
Household Income under 50k

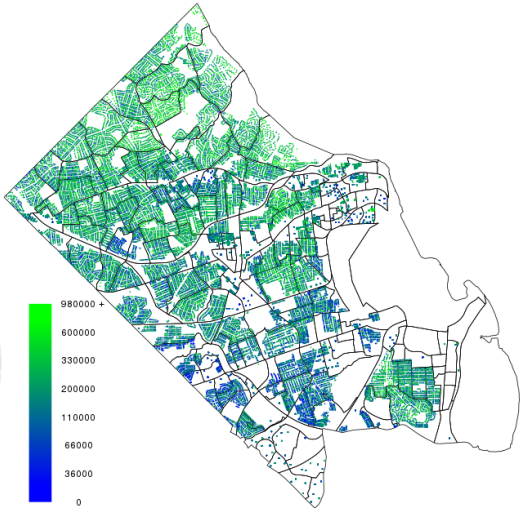
Fairfax County (2011-2015, 5-year estimate)

Example: Inferring household attributes in Arlington









Local Data
housing price
tax amount from
real estate tax
assessments

ACS
Income summary
- by block group

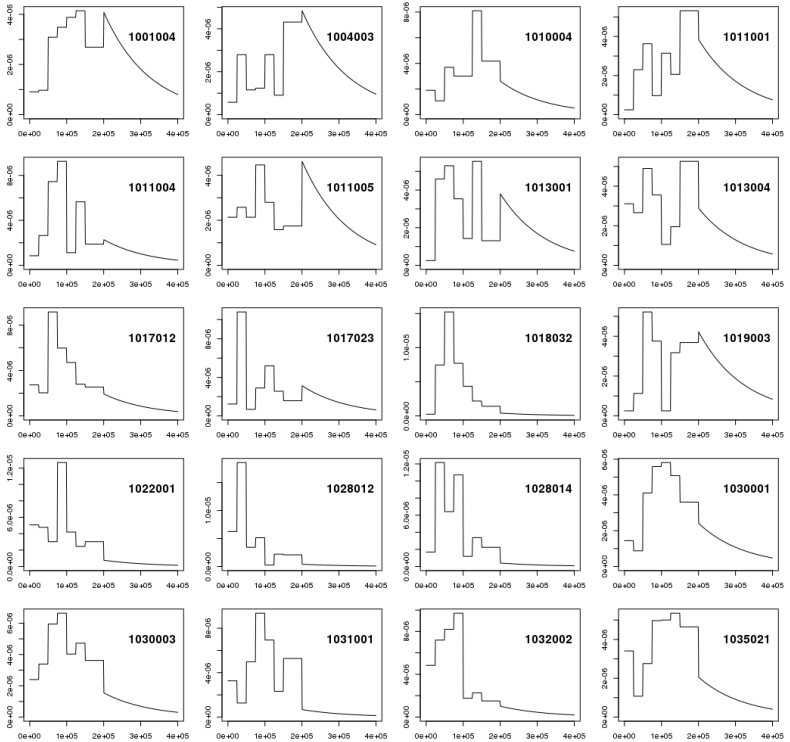
ACS Microdata
housing price
tax amount
income
- county sample

⇨

Household level Imputation
housing price
tax amount
income

ACS Summary Tables by Block Group

Use summary counts from American FactFinder to create marginal distributions for a variable (e.g. household income) in each block group.

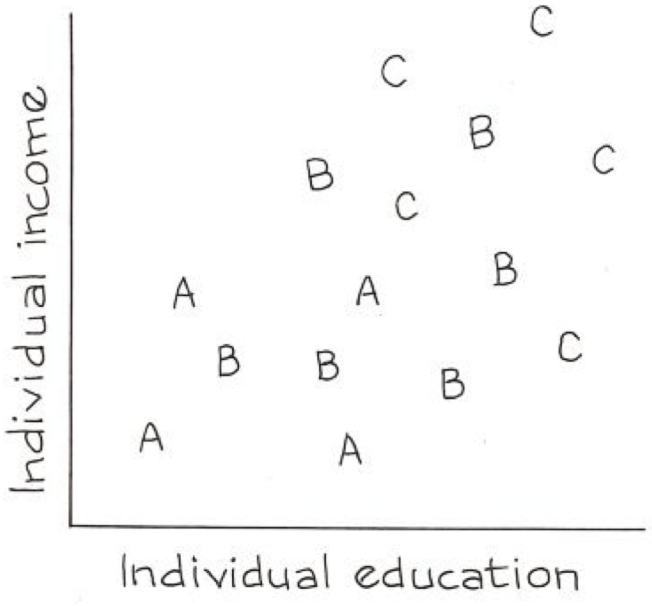


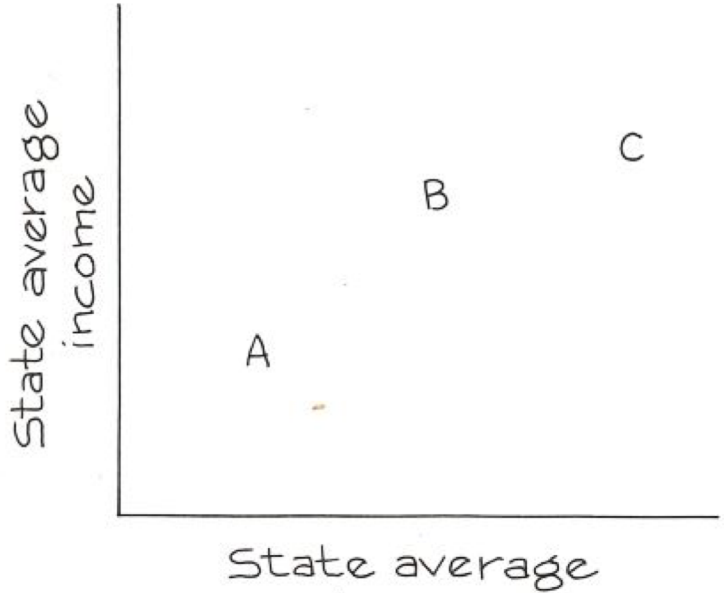
Results from identifiability

- some realizations of Arlington
  - income for each house location jg working on it!
- identifiability
  - histogram of # of unique houses in Arlington jg working on it!
  - histogram of # of unique records in a SRS of size HUD survey jg working on it!
  - CI's for probability of a unique household for different block groups (same as picture?)
  - Medians plot by block group
  - expected number of identifiable units by block group map. jg working on it!
- Be sure I list variables that are being "imputed" in the HUD sample

Individual inferences are difficult with aggregated data

- Relationships of aggregates typically overstate the strength of the relationship.
- Aggregates do not show relationships at the individual level
- Data are commonly released giving information aggregated over spatial regions
- Different data sources often use different aggregation units (e.g. school districts vs counties)
- Health and policy actions take place at the individual level



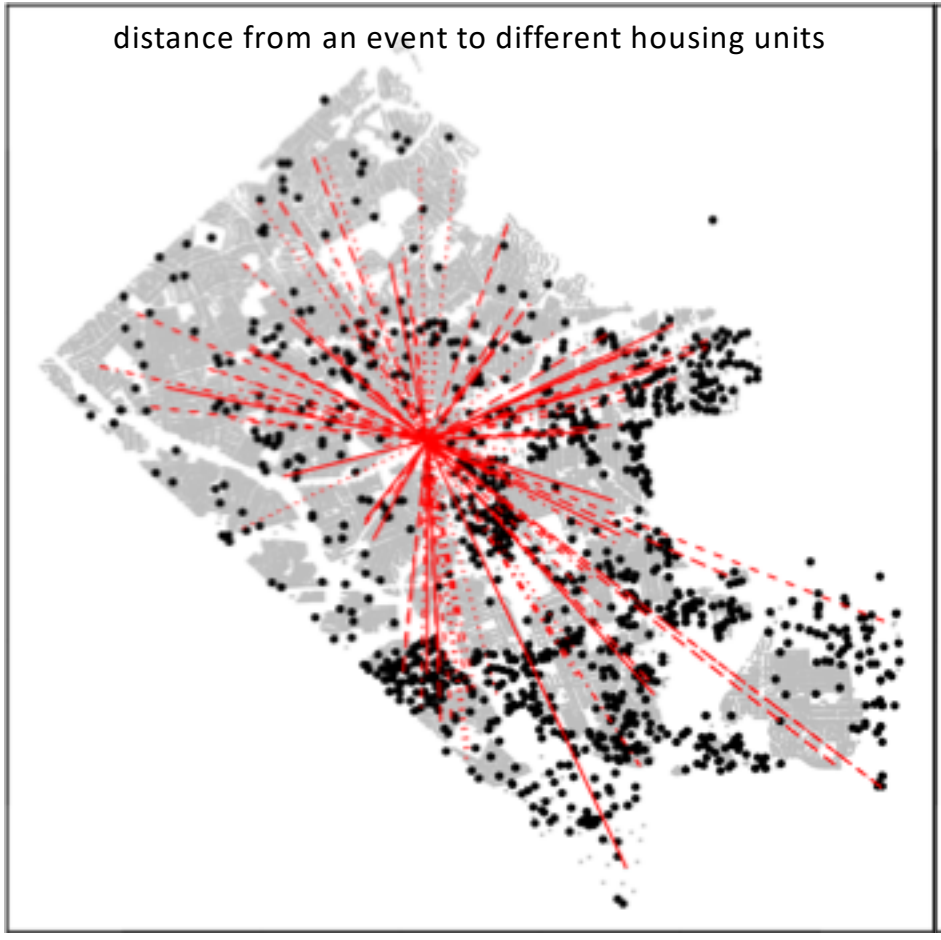


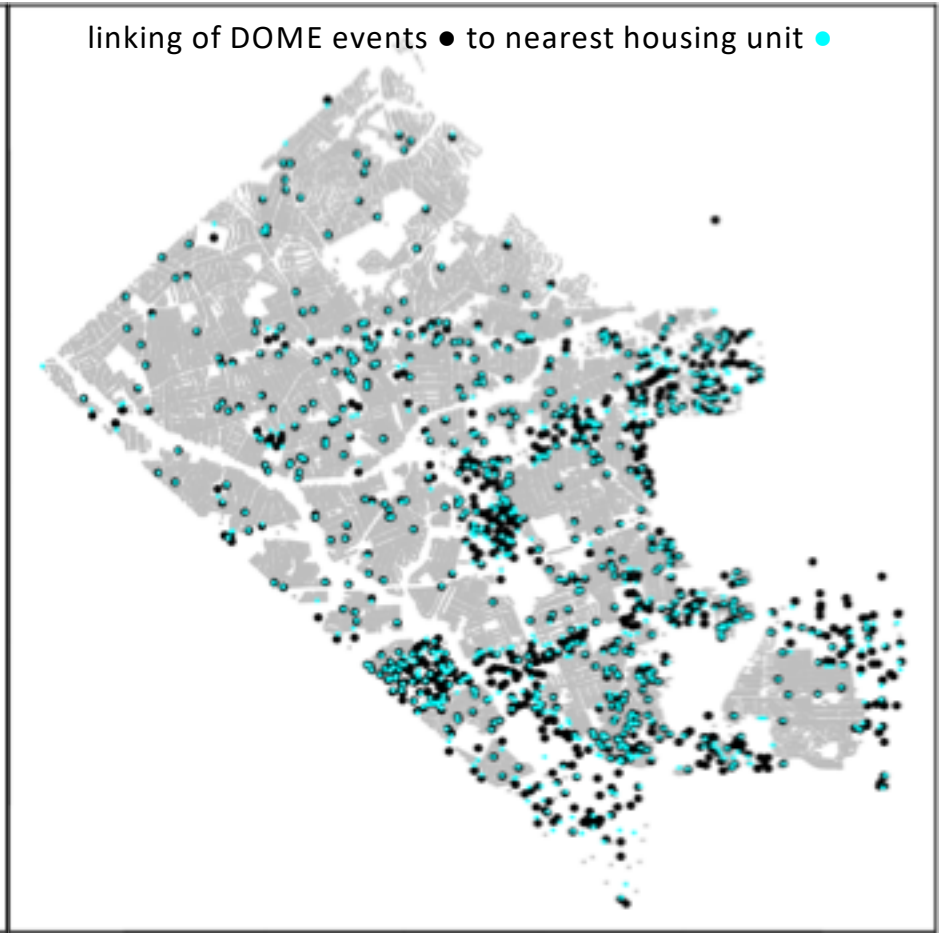
From *Statistics* by Freedman, Pisani & Purves

Methodology Update

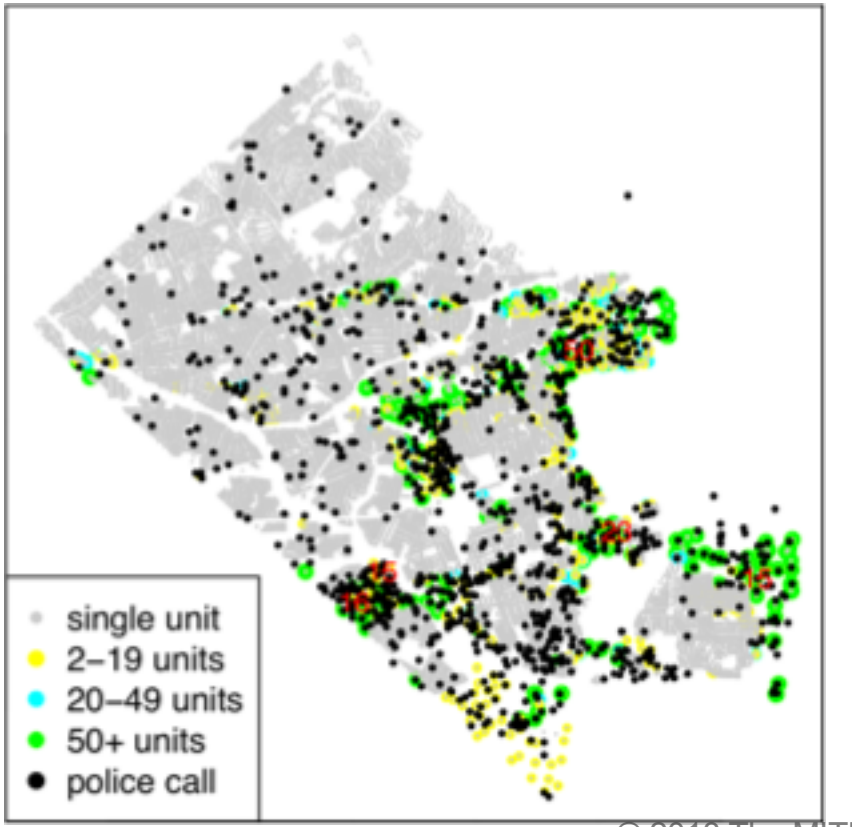
- Initial Synthetic Arlington realizations online
  - Using just 3 variables for now (housing price, household income)
  - Working to include apartments now
    - Clearly an important collection of housing units for domestic abuse
    - Most common police calls for DOME are at large apartment buildings
      - Not clear if the rate is higher since the number of units is higher.
- Police call locations for DOME are currently being used as spatially located events for analysis
- Initial approaches to link event locations to household locations
  - Nearest household location to DOME event is currently being used.
  - Will consider more model based approaches for making this link
    - Approaches will adapt "smart scatter" methods producing domestic abuse events that are consistent with police records and aggregated counts from other sources (e.g. calls to Arlington CPS by zipcode).
- Work is ongoing to include additional demographic variables.

Initial linking based on distance to DOME event



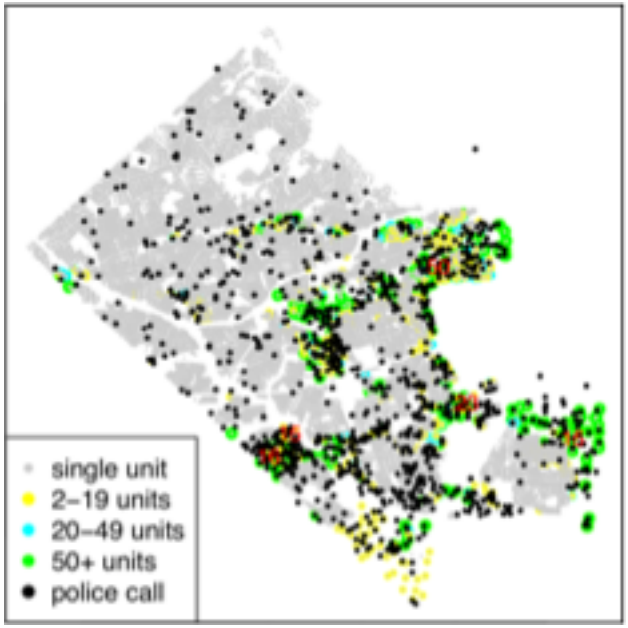


Arlington Housing Units and Arlington County Police Department (ACPD) Calls for Domestic Violence

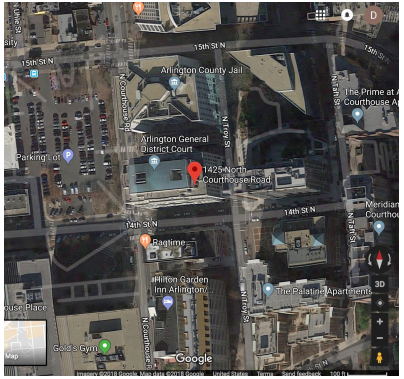


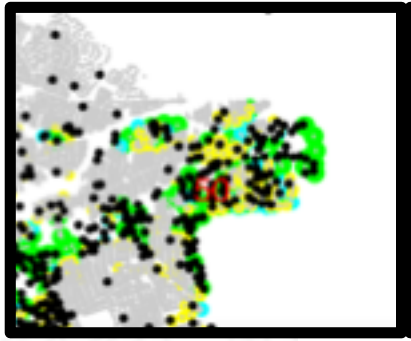
- Spatial patterns in locations for domestic dispute (DOME) calls to ACPD
- Housing types also show a similar spatial pattern
- Higher density of incidents corresponds to higher density of housing units
- Housing data allows for local estimation of rates of DOME calls

Arlington County Police Department Call for Service Data: Zooming in on High Count Repeat Locations

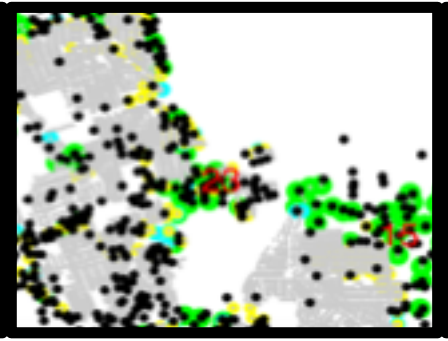


1425 N COURTHOUSE RD

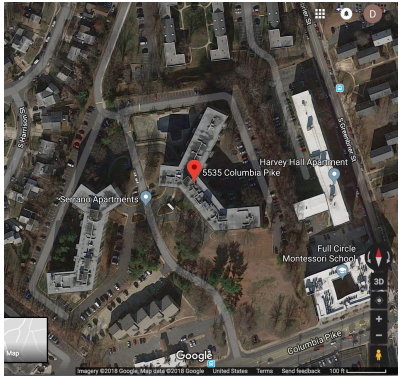


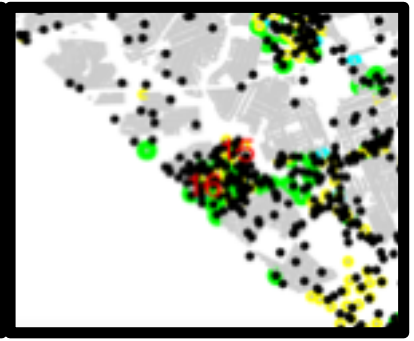


1850 COLUMBIA PIKE

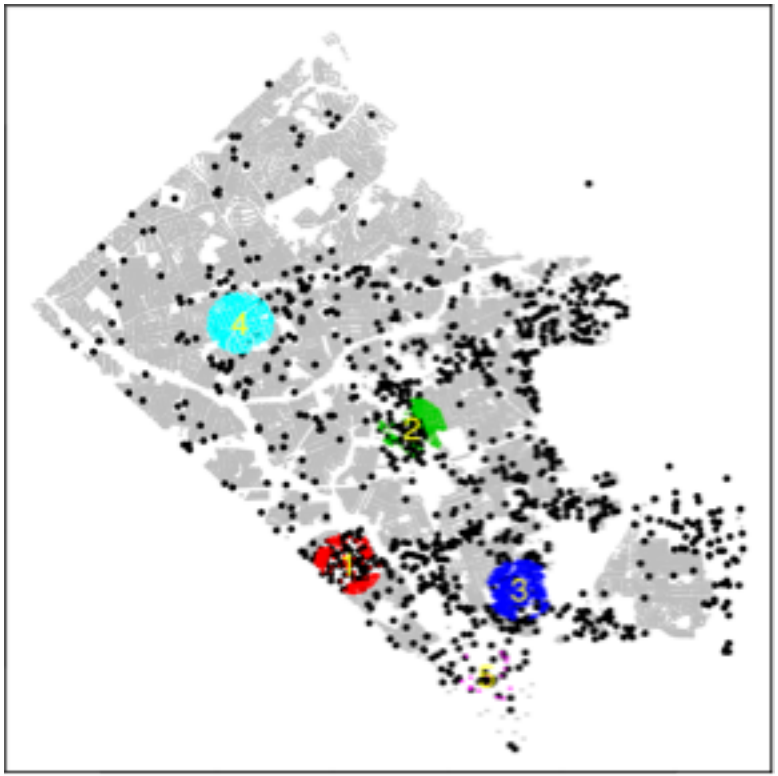


5535 COLUMBIA PIKE





Local Data Allows Estimation of Local Domestic Dispute Call Rates

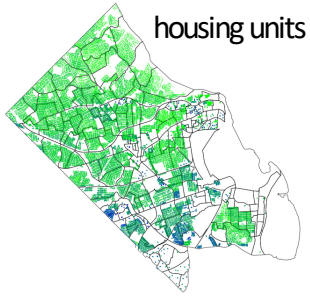


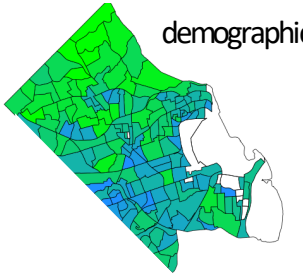
- Housing data allows for local estimation of rates of DOME calls
  - local tax records and apartment data allow estimation of the number of households in a region
  - half km radius about each of 5 points
  - Rates (# DOME events/half-km radius) for the 5 regions:

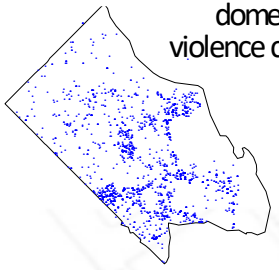
1 (red)	2 (green)	3 (dark blue)	4 (teal)	5 (pink)
0.001	0.000	0.011	0.015	0.002

Smart Scatter Model Challenges: Constructing Synthetic Arlingtons

Multiple data sources at various levels of aggregation.







Combine using state of the art concepts from:

- Multiple Imputation
- Synthetic population generation
- Bayesian Computation

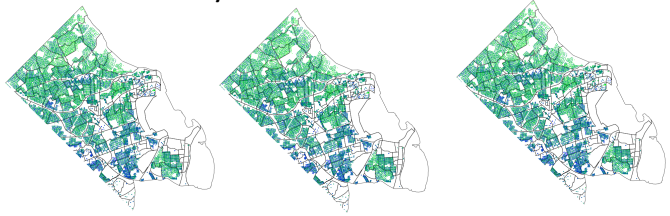
Generate plausible fine scale configurations of the population/system for prediction and quantifying uncertainty.

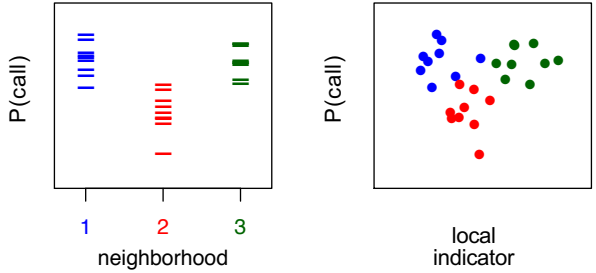
Local data sources relevant to understanding factors relating to domestic abuse calls:

- Real estate tax records
- Demographics (e.g. American Community Survey)
- Local Police Responses
- Local EMS Medical Responses

Posterior Realizations

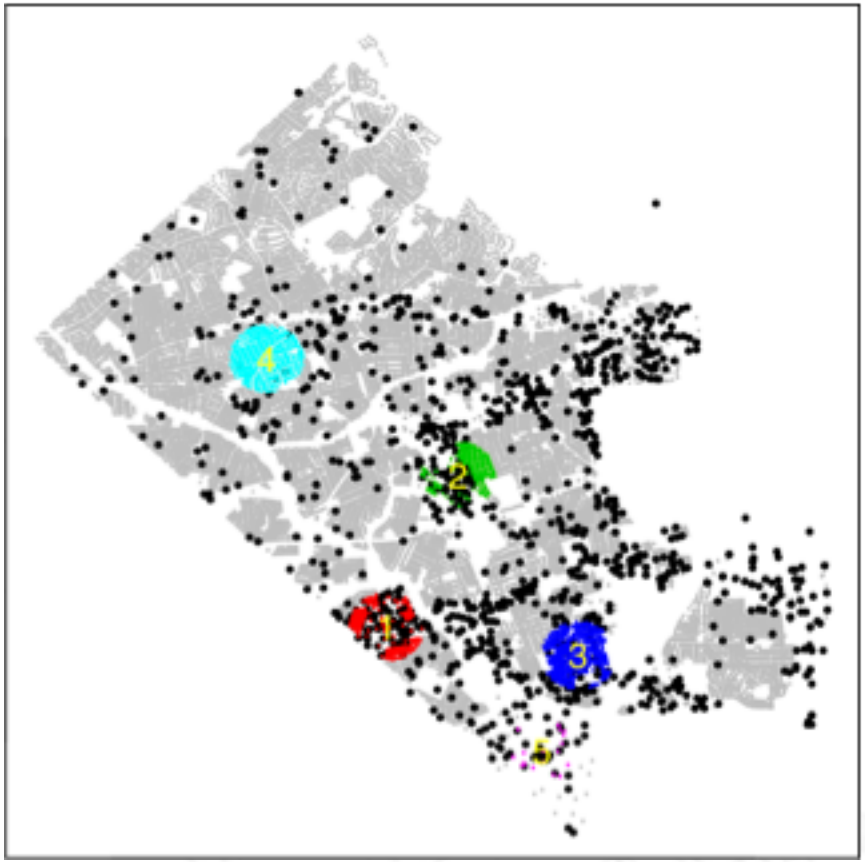
- incorporate information and uncertainty in data combination

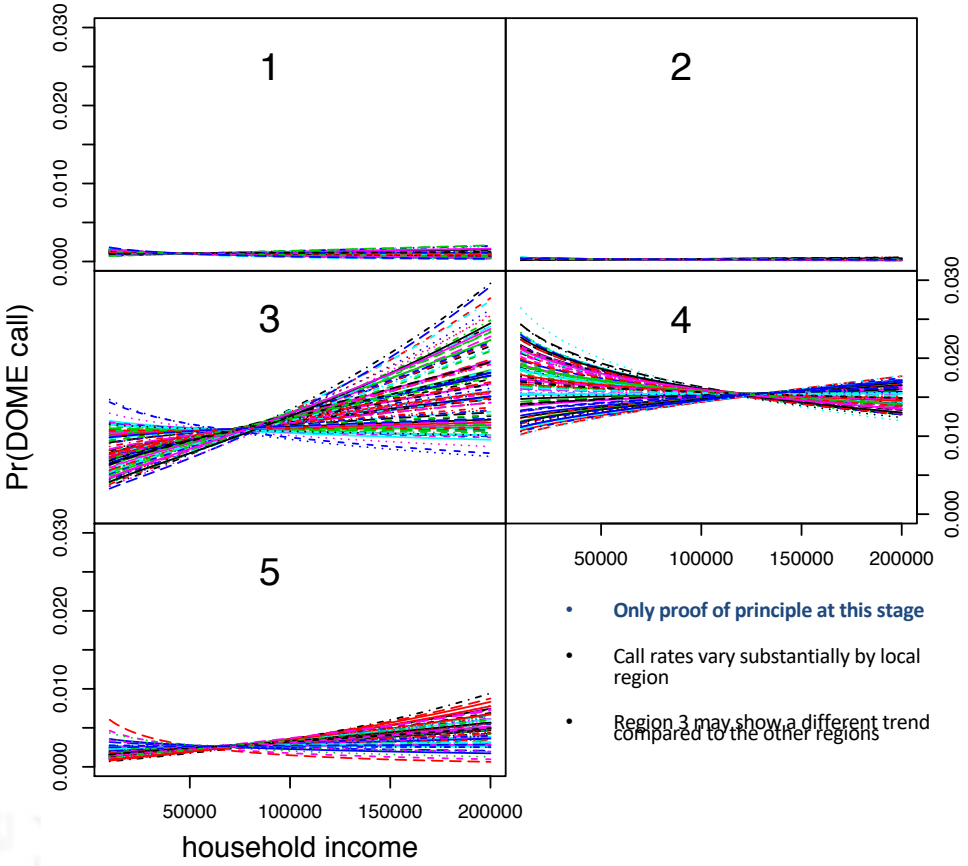




Statistically generated realizations allow separation of signal and noise at the local level.

Local Association between Domestic Dispute Call Rates and Income

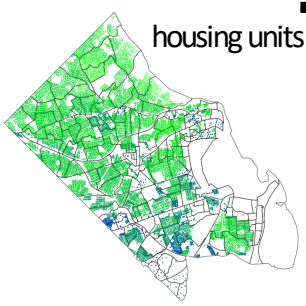


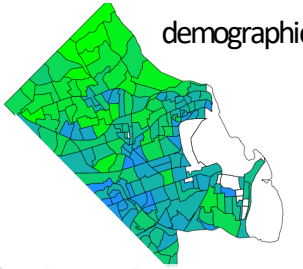


- **Only proof of principle at this stage**
- Call rates vary substantially by local region
- Region 3 may show a different trend compared to the other regions

Combining local data sources with a probabilistic representation of the local community

Multiple data sources at various levels of aggregation.





Combine using state of the art concepts from:

- Multiple Imputation
- Synthetic population generation
- Bayesian Computation

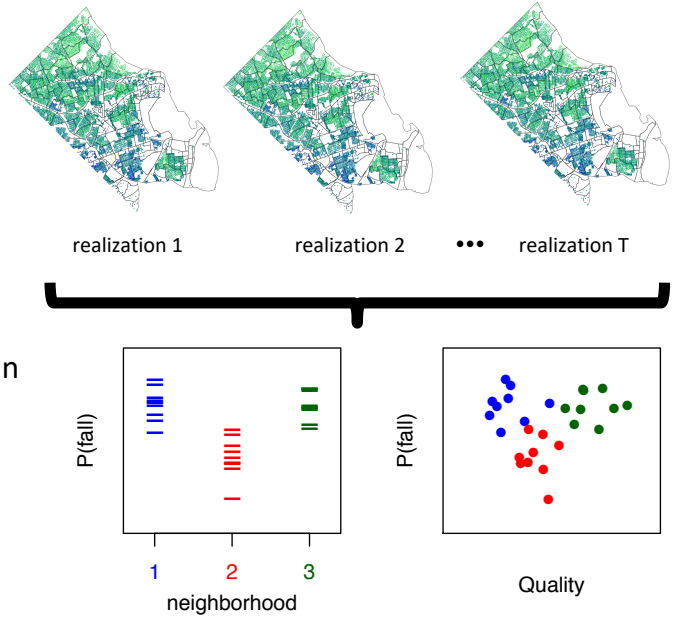
Generate plausible fine scale configurations of the population/system for prediction and quantifying uncertainty.

Local data sources relevant to home and recreation safety among elderly include:

- Real estate tax records
- Demographics (e.g. American Community Survey)
- Local EMS Medical Responses
- Location and usage of recreation facilities

Posterior Realizations

- incorporate information and uncertainty in data combination



Statistically generated realizations allow separation of signal and noise at the local level.

Imputation Scheme

Multivariate Imputation by Chained Equations (mice)

- Each variable with missing data is modeled conditionally on other variables
- Impute from each conditional distribution using Markov chain Monte Carlo methods; iterate until convergence
- Advantages:
  - Natural framework for combining multiple partially observed data
  - Imputed draws are unique and not resampled from the PUMS (as in iterative proportional fitting)

van Buuren and Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R."

Details of Demo example

- house value, property tax, location from tax records
- house value, property tax and HH income from ACS
- income distribution by block group
- ➔ produce realizations of Arlington with incomes for each household that are compatible with the marginal income tables, the tax records, and the microdata.