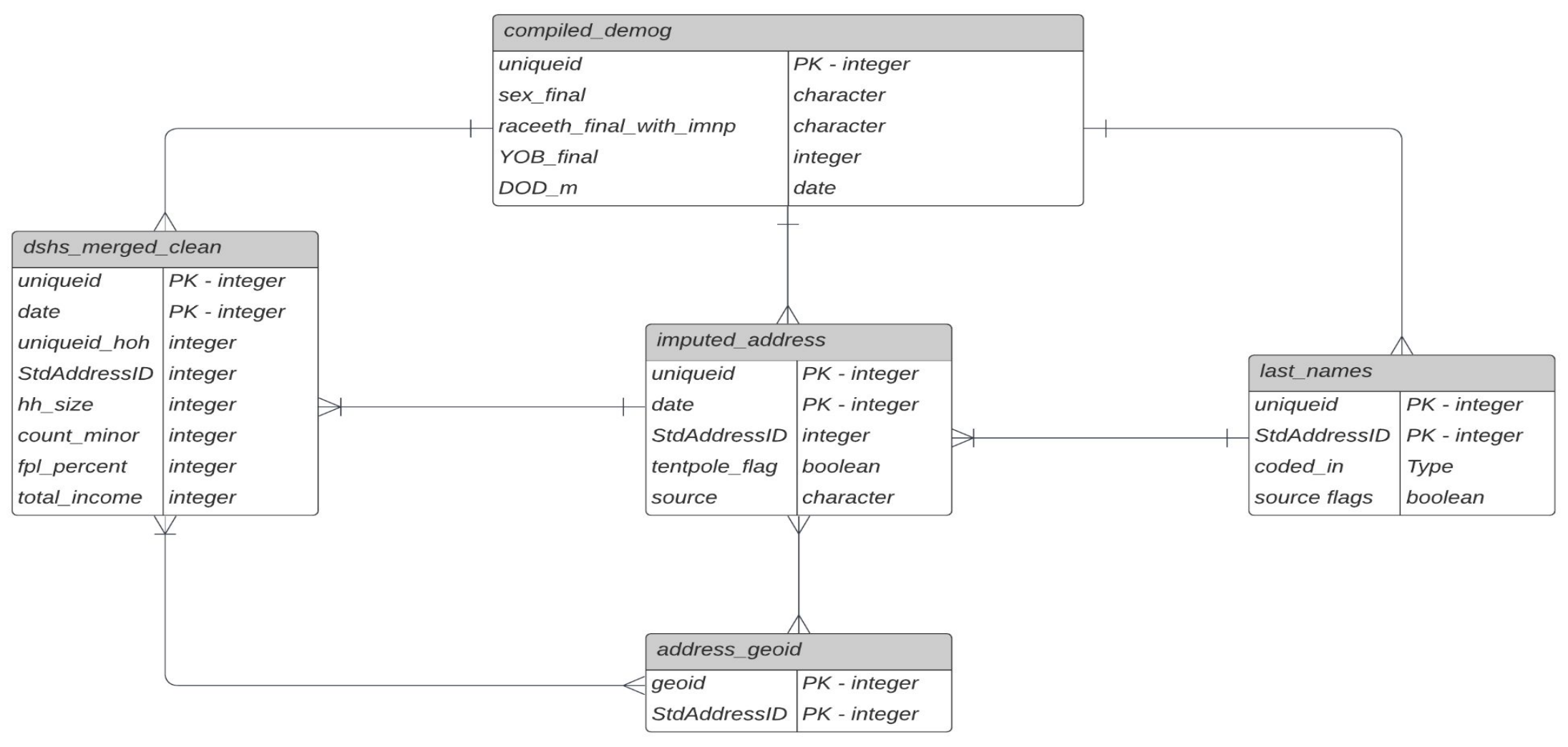
Building Households and Families out of Individual-Level Administrative Data

Zhaowen Guo and Eliot Stanton

ABSTRACT

Administrative data is collected at an individual level, but poverty and many other social outcomes are measured on a household level. This research project focused on grouping individuals into households and families out of the Washington Merged Longitudinal Administrative Dataset (WMLAD), a compilation of administrative records from six state agencies. Using point-in-time and longitudinal approaches on address and last name data, we designed and implemented four different definitions of a household or family unit to capture a variety of household types reflecting our complex society. The definitions, code, and resulting household groupings will be used for future research at a household level. Additionally, this project restructured and organized relevant data from WMLAD into a relational database to make this administrative data user-friendly for future researchers.

DATA



compiled_demog

uniqueid	PK - integer
sex_final	character
raceeth_final_with_imnp	character
YOB_final	integer
DOD_m	date

dshs_merged_clean

uniqueid	PK - integer
date	PK - integer
uniqueid_hoh	integer
StdAddressID	integer
hh_size	integer
count_minor	integer
fpl_percent	integer
total_income	integer

imputed_address

uniqueid	PK - integer
date	PK - integer
StdAddressID	integer
tentpole_flag	boolean
source	character

last_names

uniqueid	PK - integer
StdAddressID	PK - integer
coded_in	Type
source flags	boolean

address_geoid

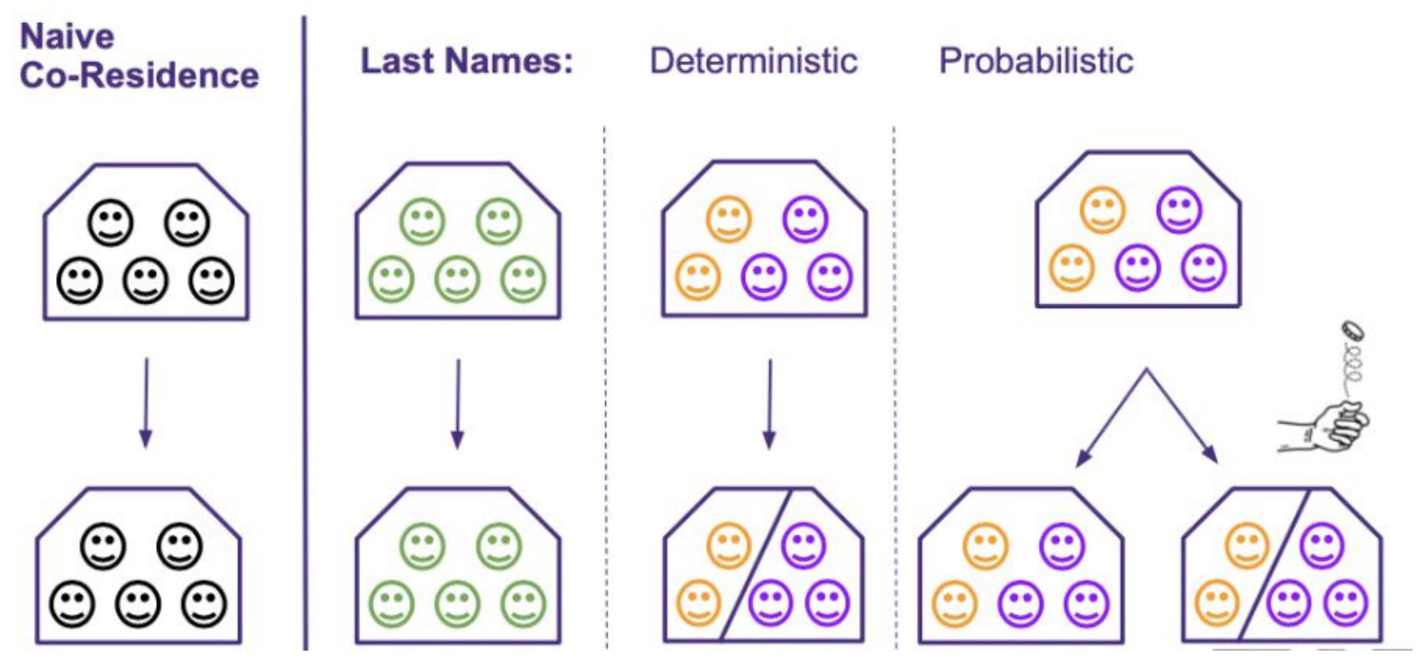
geoid	PK - integer
StdAddressID	PK - integer

> We cleaned the data of interest and built a PostgreSQL database to clarify variable meanings and make querying and restructuring the data less computationally-expensive.

> We worked primarily with addresses, most of which were imputed by other WMLAD researchers to fill in the temporal gaps between sporadic address records.

POINT-IN-TIME APPROACH

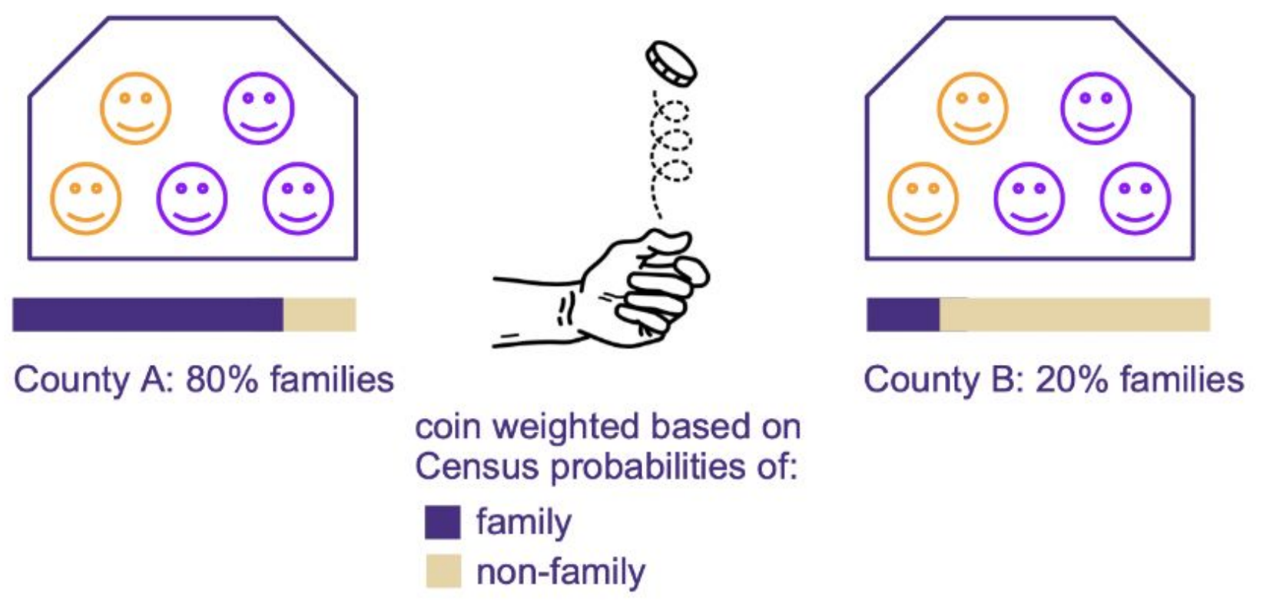
We created three different point-in-time household definitions, each using only a single month of data.



> Naive Co-Residence
The naive coresidence definition groups the entire population by their imputed address for a given month and labels them as a household. This definition fails to capture whether people who live together are roommates or family.

> Last Name (Deterministic)
For the last name deterministic definition, we keep the naive definition for the residences where everyone shares a last name. When there are multiple last names, we split the residence into multiple households, one for each last name.

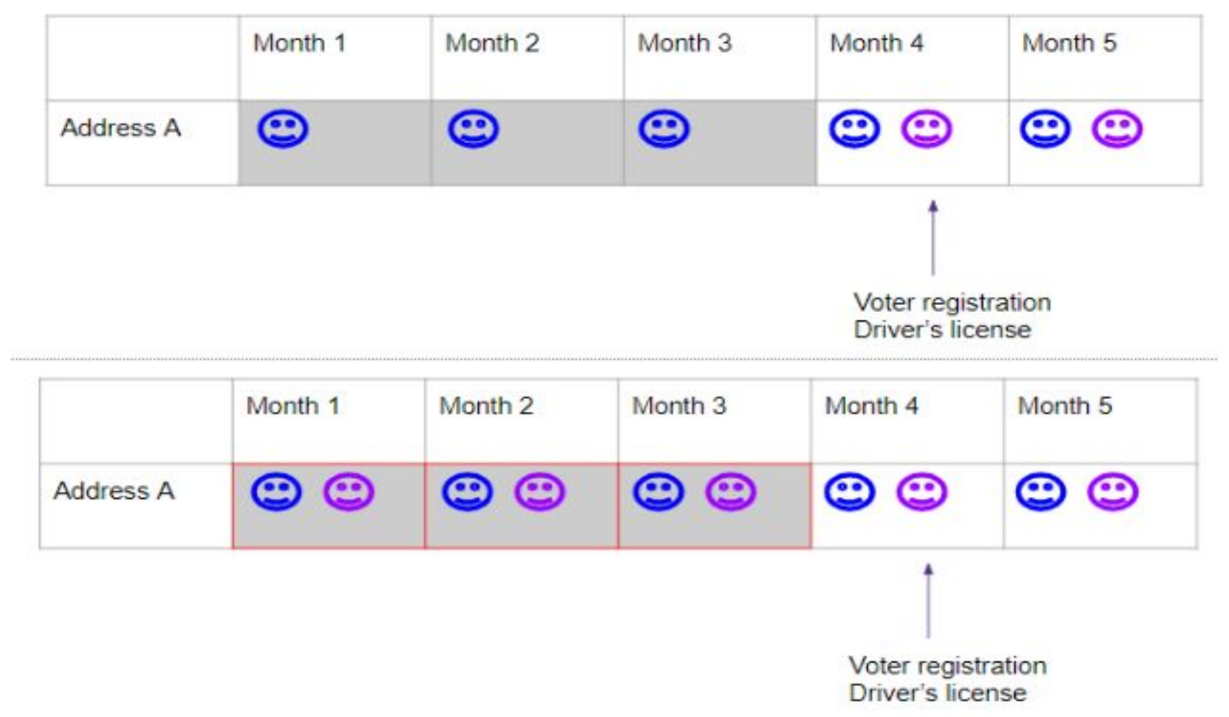
> Last Name (Probabilistic)
Because we know that many families contain multiple last names, we try using a weighted coin for the residences with multiple last names to decide if we should keep the naive definition or split the people apart by last name. The coin is weighted by the likelihood that a residence is a family household according to Census data about household size and geographical tract.



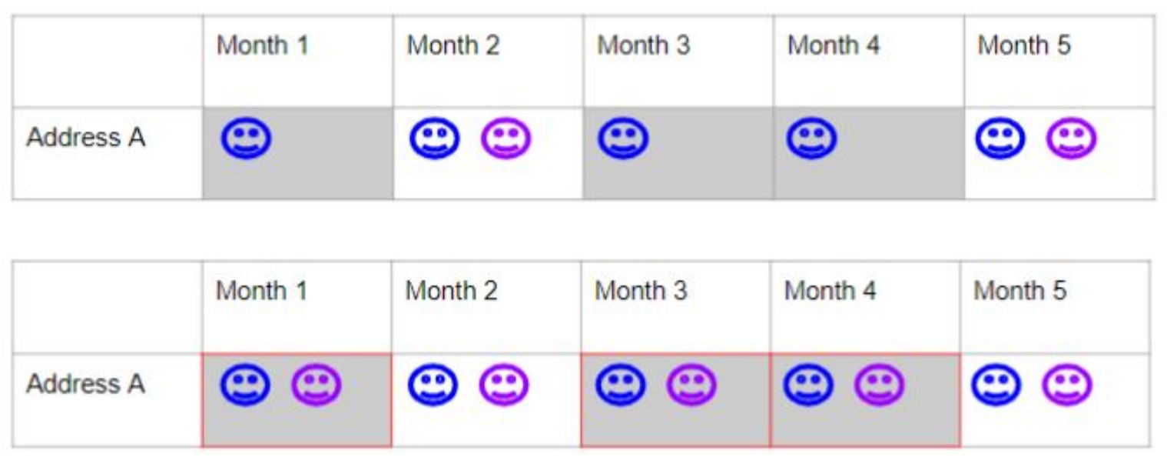
LONGITUDINAL APPROACH

By comparing the naive co-residence from WMLAD and Census 2010, we found that we overestimated one-person residences by 11%. Why might this be, and how can we adjust the overestimation?

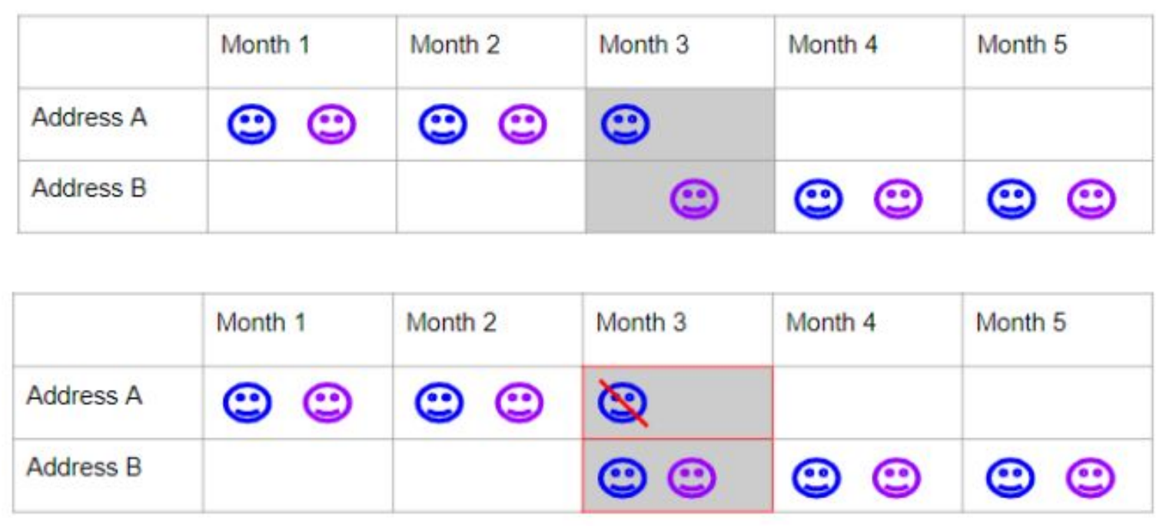
> Lack of interactions
Some household members, especially children, interact with government agencies less frequently than adults, so they may seem to appear in a household at the time of their first recorded address.



> Imputation lag
Frequent movements into and out of the same place are likely to represent an imputation error.



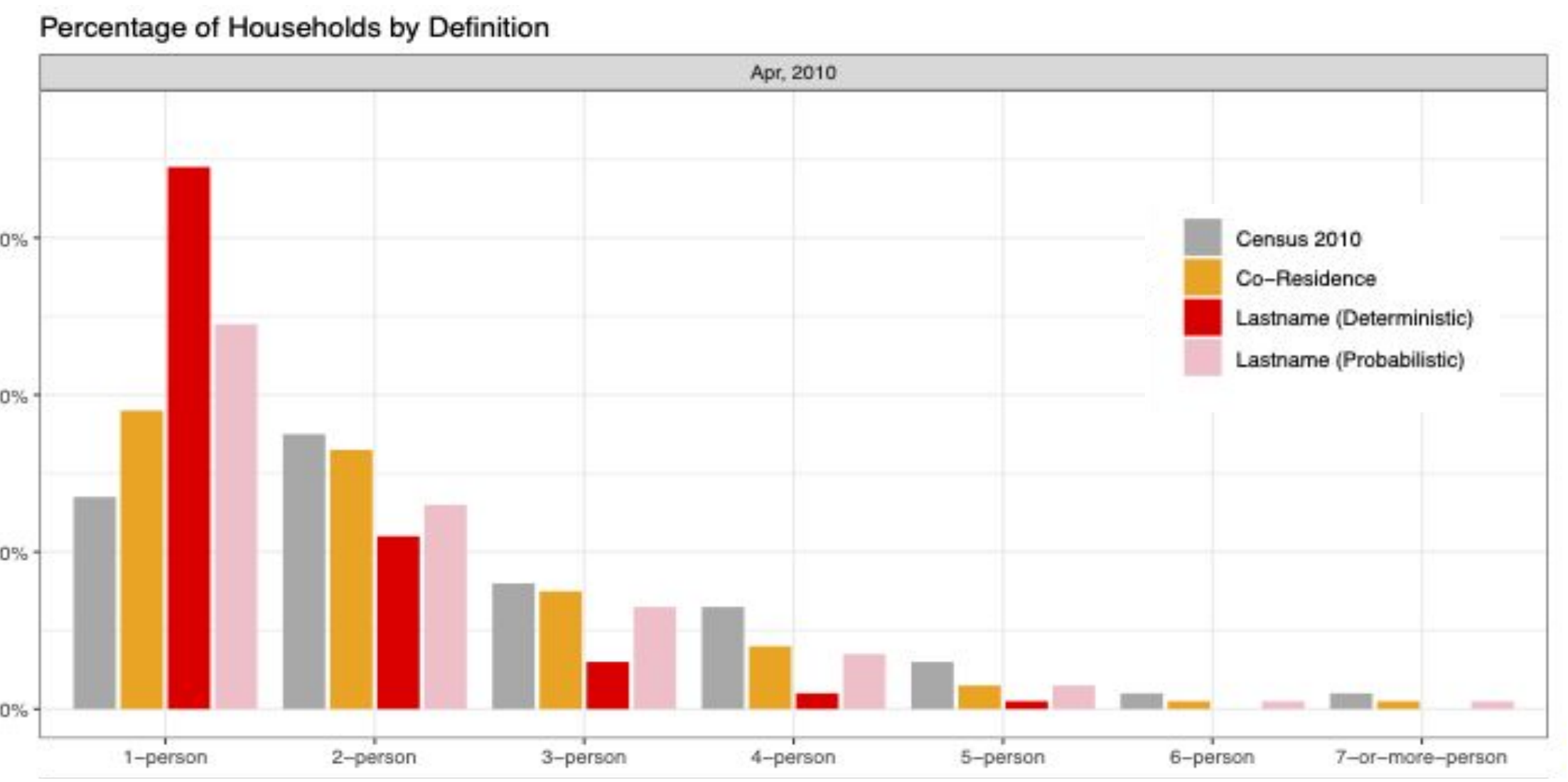
> Move in and out
One-person residences could also occur in the transition period when a multi-person household moves together to a new address but their addresses are not updated at the same time.



After revising one-person residences that fall under any of the scenarios, we leveraged recorded addresses to remove duplicated residents, making sure that each person could only occur once in a given month.

RESULTS

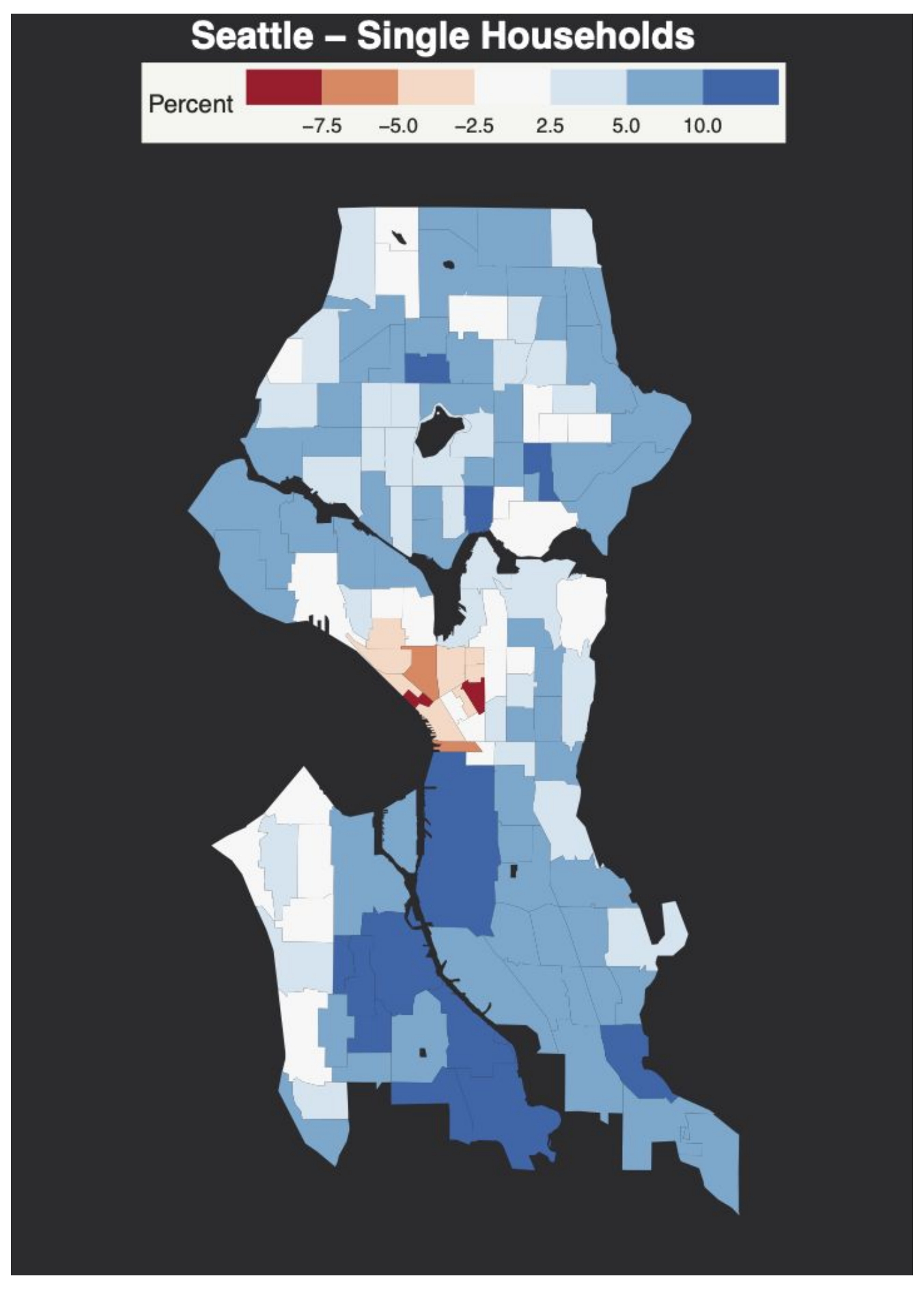
> Point-in-time approach
The naive co-residence method closely mimics the Census distribution of household sizes. As expected, the last-name-based definitions measure a concept closer to family and therefore show more one-person households, as each residential address may contain multiple unrelated individuals.



> Longitudinal approach
After implementing the modification algorithm, the proportion of one-person residences decreased from 38% to 31%, becoming closer to the census data.

	Census	Naive co-residence	Modified co-residence
Household size	Percentage	Percentage	Percentage
1	0.27	0.38	0.31
>1	0.73	0.62	0.69

> Spatial variation
By visualizing the difference in the percentages of one-person households between WMLAD co-residence and the 2010 Census, we found that we underestimated single households in downtown Seattle while overestimating them in other areas.



FUTURE WORK

> Apply the longitudinal approach to households of larger sizes

> Integrate point-in-time family groupings and longitudinal corrections

> Incorporate other information (i.e. social networks, demographics, anti-poverty programs) to capture a diversity of households