DEEP-MAPS Model of the Labor Force

Yair Ghitza*
Chief Scientist, Catalist

Mark Steitz
Principal, TSD Communications
Adjunct Professor, Columbia University

August 4, 2020

WORKING PAPER / DRAFT

Abstract

We introduce the DEEP-MAPS model of the labor force—*Demographic Estimates of Employment and Participation, with Multistage Adjustment and Poststratification Synthetics*. The model projects official government labor force statistics to small subgroups of the population, enabling granular analysis of the labor force, conditional on the model's assumptions. It works by using a modification of multilevel regression and poststratification (MRP) to estimate the relationship between demographics, geography, and labor force statistics in the Current Population Survey (CPS), paying particular attention to demographically adjusted geographic predictors of industry, occupation, and recent employment status. These relationships are projected to a synthetically built, full national joint distribution at state, county, and census tract levels. Multiple stages of adjustment are used to ensure close proximity to official sources like the Local Area Unemployment Statistics (LAUS) division and the official CPS numbers. The model is currently implemented down to the census-tract-by-demographic level, producing a complete national database of inferred labor force participation, employment, and "at work" status (without seasonal adjustment). The model uses 100% publicly available government data, and code will be made available to expedite analysis of the labor force in response to COVID-19 and model improvement. In this paper, we describe the model's assumptions and steps in detail, and we provide examples of how it can be used to facilitate deep examination and analysis of the labor force.

*We thank Lewis Alexander, Robert Dent, Andrew Gelman, Jed Kolko, Michael Mandler, Jonathan Robinson, David Rothschild, Michael Stepner, and Paul Wolfson for helpful comments on earlier versions of this paper. All errors are our own.

1 Introduction

The unprecedented, fast changing, and disparate impacts of COVID-19 present profound challenges to political leaders, policy makers, businesses, and citizens. There is considerable interest in the economic impact of the current crisis, especially what has happened to the labor force and unemployment in different communities across the United States.

Employment statistics from the Current Population Survey have been the work horse of labor market analyses since March of 1940 when the survey was initiated as a part of the Works Progress Administration[1]. Researchers have noted two limitations with the CPS data in terms of the pandemic:

> First, such data are often available only with a significant time lag. For instance, the Employment Situation Summary (i.e., jobs report) released by the Bureau of Labor Statistics on May 8 presents information on employment rates as of the week ending April 12; the next update will not come for another month. Second, due to limitations in sample sizes, such statistics typically cannot be used to assess granular variation across geographies or subgroups; most statistics are typically reported only at the state level and breakdowns by demographic subgroups or sectors are often unavailable.[2]

Many researchers are working to develop reliable, timely, and granular information about the labor force by using non-CPS data sources that offer more timely insights. These include anonymized administrative payroll data to estimate employment impacts[3], scheduling software reports for small and medium sized businesses[4], vacancy listings and unemployment claims[5], original survey data[6], and platforms that combine many sources into a single place[7]. These efforts have each made contributions to timely understanding of the differential impact of the crisis on the labor market.

This paper seeks to add to this work by utilizing the CPS microdata and modern statistical analysis techniques that directly address the ability to provide more granular estimates of the impact. These estimates do not help on timeliness, but it is hoped they may offer added insight into the geographic and demographic impacts of COVID-19.

An overview of our method can be seen in Figure 1 and is described in detail in Section 2. We combine data from three government sources—the Current Population Survey (CPS), the Local Area Unemployment Statistics (LAUS) program, and the American Community Survey (ACS)—leveraging the unique strengths of each. The CPS provides high quality labor force statistics for the nation as a whole and for demographic groups at a national level; the LAUS provides topline estimates of labor force participation and unemployment for smaller geographies; and the ACS provides detailed, large-scale data on the demographic distribution of the country.

[1] John E Bregger. "The Current Population Survey: A Historical Perspecitve and BLS Role". In: *Monthly Lab. Rev.* 107 (1984), p. 8; Megan Dunn, Steven E Haugen, and Janie-Lynn Kang. "The Current Population Survey: Tracking Unemployment in the United States for Over 75 Years". In: *Monthly Labor Review* (2018), pp. 1–23.

[2] Raj Chetty et al. "Real-Time Economics: A New Platform to Track the Impacts of COVID-19 on People, Businesses, and Communities Using Private Sector Data". In: (2020).

[3] Tomaz Cajner et al. "The US Labor Market During the Beginning of the Pandemic Recession". In: (2020).

[4] Alexander W Bartik et al. "Labor Market Impacts of COVID-19 on Hourly Workers in Small-And Medium-Sized Businesses: Four Facts From Homebase Data". In: (2020); Andre Kurmann, Etienne Lale, and Lien Ta. "The Impact of COVID-19 on US Employment and Hours: Real-Time Estimates With Homebase Data". In: *Unpublished Manuscript* (2020).

[5] Lisa B Kahn, Fabian Lange, and David G Wiczer. "Labor Demand in the Time of COVID-19: Evidence From Vacancy Postings and UI Claims". In: (2020).

[6] Ernie Tedeschi and Quoctrung Bui. "America's Employment Losses Might Be Slowing: Job Tracker". In: *The New York Times* (2020); Alexander Bick and Adam Blandin. "Real Time Labor Market Estimates During the 2020 Coronavirus Outbreak". In: *Unpublished Manuscript, Arizona State University* (2020); Olivier Coibion, Yuriy Gorodnichenko, and Michael Weber. "Labor Markets During the Covid-19 Crisis: A Preliminary View". In: (2020).

[7] Raj Chetty et al. "Real-Time Economics: A New Platform to Track the Impacts of COVID-19 on People, Businesses, and Communities Using Private Sector Data". In: (2020).

Using the CPS microdata, we fit flexible regularized statistical models of various aspects of the labor force, estimating their rates conditional on demographics, state, and *demographically adjusted geographic predictors* (DAGPs): historical unemployment, labor force participation, and industry/occupation trends[8]. These DAGPs can be thought of as ACS/LAUS estimates at the county or census tract level, where each value has been allocated to the different demographic groups inside of each geography[9]. The models are projected to a detailed poststratification dataset, representing the full joint distribution of relevant variables. Joint demographic distributions can be produced fairly easily at the state level using ACS microdata, but they are generally unavailable for counties and census tracts. We create that data synthetically, combining marginal or semi-joint distributions for each geography, modeled conditional relationships from the large-scale ACS microdata, and modest geographic smoothing across census tracts[10]. This level of projection creates somewhat noisy estimates, which we adjust through various levels of correction: at the county, state, demographic, and national level[11]. We summarize this as the DEEP-MAPS model: *Demographic Estimates of Employment and Participation, with Multistage Adjustment and Poststratification Synthetics*.

Conceptually, one way to think about our approach is as follows: we start with last month's county-level LAUS numbers for labor force participation and unemployment and adjust them based on demographics and geographic industry/occupation trends, all guided through a flexible model fit to the CPS microdata. We project that model to a detailed demographic database of every census tract in the country, built synthetically from the ACS, and then constrain the resulting estimates to guarantee they are very close to the national CPS data and the LAUS state/county estimates (without seasonal adjustment)[12]. The value comes from the newly available granularity, which are available down to the census tract level and for demographics within each geography.

This paper describes this method in detail, and provides examples of what can be learned from the resulting output. In order to make our methods transparent and allow other researchers to extend and improve upon them, we exclusively use publicly available data, and we will be making our code fully available[13].

While we think our estimates add value, we urge caution on various fronts. First, our estimates are *modeled inferences*, not *reported data*. They should be seen as suggestive rather than definitive evidence, with uncertainty around them, especially for the smallest groups. This caveat may be obvious for experienced modelers, but the allure of "big data" and sophisticated methods often confuse people into assigning too much specificity to modeled inferences. Second, the impact of COVID-19 on the labor force is an incredibly fast-moving situation, accompanied by many types of data and reporting problems, including delayed reporting, unusual patterns of survey non-response, survey misreporting, and other issues. As such, we interpret our estimates, as well as data from other sources, as having a larger amount of uncertainty around them than usual. We believe analysts and policy-makers should not exclusively rely on any single piece of evidence, and would be better served by comparing different sources to examine consistency and robustness. Our hope is that this paper contributes to that endeavor.

[8] Section 2.2
[9] Section 2.1.2
[10] Section 2.1.1
[11] Section 2.3
[12] We understand that both the CPS and the LAUS are estimates themselves, which can have problems due to survey non-response, modeling assumptions, and more. However, the methods used at the BLS have been built up over many years of careful work, and we do not attempt to improve upon them here. Further, we avoid seasonal adjustment for two reasons. First, seasonal adjustment is intended to control for regular fluctuations of the business cycle; in our view, COVID-19 is such a large shock that understanding how it integrates with the regular business cycle is challenging to say the least. Second, to be frank, seasonal adjustment is complicated, and determining the best way to project it to small geographic and demographic subgroups is outside the scope of this paper.
[13] Code can be found here.

2 Data and Methods

Our method uses publicly available data from three government sources, which complement one another in the goal of producing *small area estimates*, i.e., labor force statistics for small geographies and demographic subgroups inside of those geographies. The three data sources are the Current Population Survey (CPS), the American Community Survey (ACS), and the Local Area Unemployment Statistics (LAUS) program[14].

The CPS household survey, collected every month, is designed to reflect the civilian non-institutionalized population (CNIP). It is the primary source of labor force statistics for the population, such as the official unemployment rate for the USA. The most visible output from the CPS is the monthly Employment Situation Summary, i.e., the jobs report, which details the unemployment rate, labor force participation rate, and other important national statistics. The CPS also publishes data from the household survey in demographic tables and in publicly available, anonymized microdata files.

The LAUS program uses the CPS and other data to produce topline labor force statistics for various geographies around the country, including states, counties, and cities with population greater than 25,000. The LAUS estimates inform the allocation of resources from federal programs as well as state and local governments. But they are not produced for all geographies, and they do not allow for demographic breakdowns within geography. The scope of the LAUS is also fairly limited: they produce estimates for labor force participation, employment, and unemployment, but not for other important labor force concepts such as part time work, reasons for unemployment, and so on.

The ACS is a large-scale general purpose survey that produces population-level estimates on a variety of demographic, social, and economic characteristics[15]. Unlike the decennial census which is only conducted once every decade, the ACS is an ongoing survey and can therefore be considered a consistently reliable source for detailed population data in the country. Like the CPS, the ACS publishes their data in a variety of formats, including anonymized microdata with privacy-preserving protections on geographic indicators[16].

We build on a statistical method called multilevel regression and poststratification (MRP), that is used widely in political science and starting to be used in other fields[17]. MRP is used to project national survey data down to the state and sub-state level, where standard design-based estimates are unreliable due to lack of sample size. MRP overcomes sample size limitations and induces stability in subgroup estimates, conditional on three key assumptions. For clarity, we'll describe an example: estimating the labor force participation rate among Hispanic men in Arizona (n=207 responses in the April 2020 CPS). First, the standard design-based estimate implicitly treats this group as completely independent of any other group; the MRP process recognizes that there are likely similarities between them and, say, Hispanic *women* in Arizona (n=226), Hispanics in New Mexico (n=592), and other groups, and tries to estimate the strength of these various similarities through the data,

[14] The CPS is collected jointly by the United States Census Bureau and the Bureau of Labor Statistics; the ACS is collected by the Census Bureau; the LAUS program is a federal-state cooperative effort, whose statistics are published by the Bureau of Labor Statistics.

[15] For more detail and a comparison of employment data between the two surveys, see Braedyn K. Kromer and David J. Howard. "Comparison of ACS and CPS Data on Employment Status". In: *Census Working Papers* SEHSD-WP2011-31 (2011).

[16] We download microdata for both the CPS and ACS from IPUMS at the University of Minnesota. Their work providing a consistent coding scheme across surveys over many years was invaluable for this project. Steven Ruggles et al. *IPUMS USA: Version 10.0*. Minneapolis, MN, 2020; Sarah Flood et al. *Integrated Public Use Microdata Series, Current Population Survey: Version 7.0*. Minneapolis, MN, 2020

[17] Yair Ghitza and Andrew Gelman. "Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups". In: *American Journal of Political Science* 57.3 (2013), pp. 762–776; Yair Ghitza and Andrew Gelman. "Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research". In: *Political Analysis* (2020), 125; Jeffrey R Lax and Justin H Phillips. "How Should We Estimate Sub-National Opinion Using MRP? Preliminary Findings and Recommendations". In: *annual Meeting of the Midwest Political Science Association, Chicago*. 2013; Devin Caughey and Christopher Warshaw. *Public Opinion in Subnational Politics*. 2019; Xingyou Zhang et al. "Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System". In: *American Journal of Epidemiology* 179.8 (2014), pp. 1025–1033.

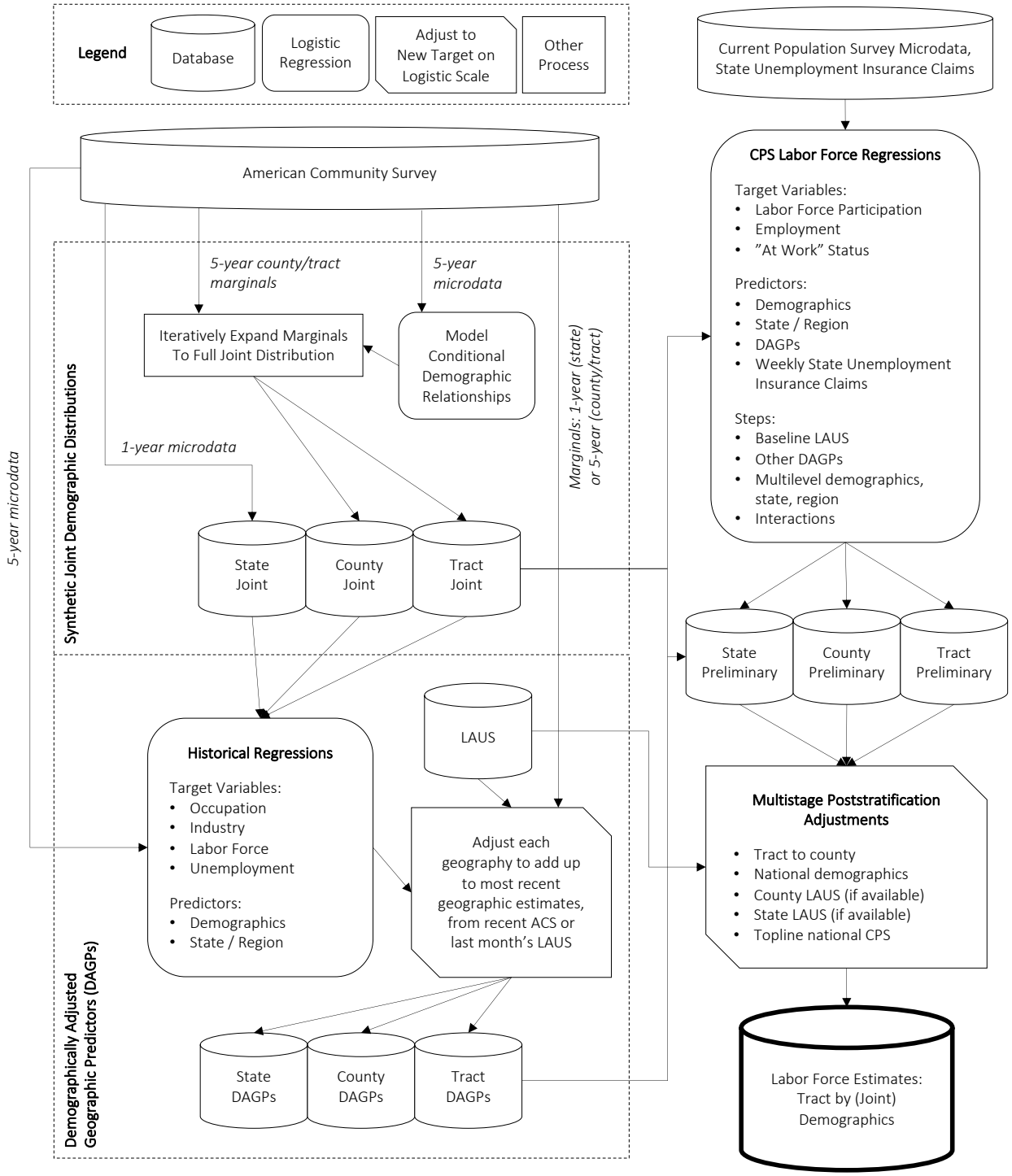

Figure 1: *Overview of data and methods.*

partially pooling responses together with a hierarchical (i.e., multilevel) regression. Second, the MRP process conditions labor force participation on the full joint demographic distribution of the group, as reported in the ACS data. For instance, the CPS shows 66% of this group as citizens with a high school degree or more, while the ACS shows that number to be 58% (n=6,486 in the most recent 2018 data). The CPS is not designed to be representative for this small of a group, so this type of disparity is to be expected. This particular difference may or may not have a big impact on the final estimate, but MRP uses the full joint data here, implicitly borrowing stability from the larger ACS data and accounting for the relationship between labor force participation and all the other covariates (education, citizenship, and others) at the same time. Lastly, MRP allows for the inclusion of auxiliary geographic data, such as the LAUS labor force participation rate from the previous month. We think of these historical geographic estimates as a starting baseline for the model, which is critical in obtaining good geographic estimates. In practice, models that do not have good geographic baselines tend to be biased toward the national average, because the included covariates do not capture the full geographic variation in the outcome variable.

This section describes the various steps in our method, detailing how we assembled the multiple necessary pieces and addressed challenges particular to this problem. Although we use this method to produce labor force estimates for earlier years, we will describe the process for April 2020 to ease the description.

2.1 Construct Poststratification Datasets

We construct datasets detailing the full joint demographic distribution of the country by three levels of geography—state, county, and census tract—for the entire citizen population over the age of 16. Once we project labor force estimates to these subgroup cells (as described in section 2.3), we can flexibly combine cells to examine labor force characteristics for any larger group of interest.

2.1.1 Demographics

Our goal is to produce a joint distribution on the demographics shown in Figure 2. This is straight forward for state-level data, because the ACS publishes large-scale microdata with state identifiers for all records[18]. Some records are identified by county, but only when that county has sufficient sample size to avoid privacy concerns. To get a national distribution at the county and census tract levels, then, we need an alternative strategy.

A recent paper suggests building *synthetic* poststratification databases in these sorts of cases, estimating joint distributions from available marginal distributions at the geographies of interest[19]. Under this approach, the marginals can be combined by either assuming conditional independence between covariates, or more accurately by using auxiliary microdata to estimate the relationship between them. As a toy example, say that we are interested in knowing the age by education distribution in a census tract:

- Groups are defined as young vs. elderly, and college vs. non-college.
- We know the marginal distributions from reported ACS data for the tract: 40% young, 60% elderly, and 35% college, 65% non-college.
- We also have auxiliary microdata from a national survey, showing that 50% of young people and 25% of elderly people are college graduates.

[18] The (most recent) 2018 ACS microdata includes 2,642,681 records among people over the age of 16. We could build a model to smooth data for smaller groups, but the sample size is sufficient to simply use the raw data here.

[19] Lucas Leemann and Fabio Wasserfallen. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation". In: *American Journal of Political Science* 61.4 (2017), pp. 1003–1022.

Demographics

Age
16-17
18-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75+

Education
No High School Degree
High School Graduate
Some College
College Graduate
Post-Graduate

Race / Ethnicity
Hispanic
Non-Hispanic:
White
Black
Asian
Native American
Other

Marital Status
Single
Married

Citizen
Yes
No

Gender
Male
Female

Demographically Adjusted Geographic Predictors (DAGPs)

Industry (ACS)
Agriculture, forestry, fishing and hunting
Mining, quarrying, and oil and gas extraction
Construction
Manufacturing
Wholesale trade
Retail trade
Transportation and warehousing
Utilities
Information
Finance and insurance
Real estate and rental and leasing
Professional, scientific, and technical services
Management of companies and enterprises
Administrative and support and waste management services
Educational services
Health care and social assistance
Arts, entertainment, and recreation
Accommodation and food services
Other services, except public administration
Public administration

Occupation (ACS)
Management, business, and financial
Computer, engineering, and science
Education, legal, community service, arts, and media
Healthcare practitioners and technical
Healthcare support
Protective service
Food preparation and serving related
Building and grounds cleaning and maintenance
Personal care and service
Sales and office
Natural resources, construction, and maintenance
Production, transportation, and material moving

LAUS (Previous Month)
Labor Force Participation
Employment Rate (As Percent of Labor Force)

Regions

State	Region
Alabama	South
Alaska	West
Arizona	West
Arkansas	South
California	West
Colorado	West
Connecticut	Northeast
Delaware	Northeast
District of Columbia	DC
Florida	South
Georgia	South
Hawaii	West
Idaho	West
Illinois	Midwest
Indiana	Midwest
Iowa	Midwest
Kansas	Midwest
Kentucky	South
Louisiana	South
Maine	Northeast
Maryland	Northeast
Massachusetts	Northeast
Michigan	Midwest
Minnesota	Midwest
Mississippi	South
Missouri	Midwest
Montana	West
Nebraska	Midwest
Nevada	West
New Hampshire	Northeast
New Jersey	Northeast
New Mexico	West
New York	Northeast
North Carolina	South
North Dakota	Midwest
Ohio	Midwest
Oklahoma	South
Oregon	West
Pennsylvania	Northeast
Rhode Island	Northeast
South Carolina	South
South Dakota	Midwest
Tennessee	South
Texas	South
Utah	West
Vermont	Northeast
Virginia	South
Washington	West
West Virginia	Northeast
Wisconsin	Midwest
Wyoming	West

Figure 2: *Covariates included in the model. Demographics are included as base terms (one-way), interactions (all two-way interactions between demographics, state, and region), and coarser groupings (i.e., 16-24 or White Non-College) to help guide partial pooling. Demographically Adjusted Geographic Predictors start with the most recent ACS or LAUS data and allocate each value to the different demographic groups inside of each geography. We also include state-level weekly unemployment insurance claims as a predictor in the primary CPS models.*

– Simple multiplication yields the joint distribution: 20% young college, 20% young non-college, 15% elderly college, 45% elderly non-college. This will be accurate if the (national) relationship found in the survey microdata holds in this census tract.

We use this concept to derive county- and tract-level joint distributions. For each county and census tract, we use marginals from the 2018 ACS summary file, and we use pooled 2014-2018 ACS microdata as the auxiliary survey data[20]. Our particular data require three adjustments to the toy example described above. First, even at the tract level, the ACS provides more than simple marginal distributions, providing all of our demographics of interest broken down by gender and age (sometimes with different age breaks across different variables). Conceptually, then, we treat a single discrete unit as reported by the ACS as a "marginal" distribution under the described framework.

Second, we are trying to estimate joint distributions for very small subgroups, where even the ACS data has small sample size[21]. As a result, when we model the conditional dependence between variables in the ACS microdata, we use a hierarchical model within each state, which provides smoothed and more reliable estimates than either the raw survey data alone or fully pooled national data[22]. For example, when we estimate the probability of being married on sex, age, race, and citizenship, the model is estimated within each state, with $y_i \in \{0, 1\}$ indicating single or married for respondent i. The model takes the following form:

$$P(y_i = 1) = \text{logit}^{-1}\left(\alpha^0 + \alpha^{gender}_{j[i]} + \alpha^{age}_{k[i]} + \alpha^{race}_{l[i]} + \alpha^{citizen}_{m[i]}\right) \tag{1}$$

$$\alpha^{gender} \sim \text{Normal}(0, \sigma^2_{gender}), \text{ for } j = \{1, 2\} \tag{2}$$

$$\alpha^{age} \sim \text{Normal}(0, \sigma^2_{age}), \text{ for } k = \{1, \ldots, 10\} \tag{3}$$

$$\alpha^{race} \sim \text{Normal}(0, \sigma^2_{race}), \text{ for } l = \{1, \ldots, 6\} \tag{4}$$

$$\alpha^{citizen} \sim \text{Normal}(0, \sigma^2_{citizen}), \text{ for } m = \{1, 2\} \tag{5}$$

where α^0 is the intercept, $j[i]$, $k[i]$, $l[i]$, and $m[i]$ refer to gender, age, race, or citizenship status for respondent i, with values following Figure 2, and the αs are varying intercepts for the same covariates. We use these probabilities to split each original cell into two—one married, one single—adding up to the same overall population. For ys with more than two classes, we fit multiple versions of these equations and ensure that the resulting probabilities add up to 1 within each cell.

Third, the multiplication in the toy example yielded a joint distribution that lined up perfectly across both marginals, which will not be the case in general. When we add a new variable (in the above case, marital status), we adjust the joint distribution to equal the marginal distribution that includes the new variable. For example, if the process above yields a within-cell distribution that is 55% married and 45% single, when the marginal distribution is 50% for each, then we multiply the married cells by 50/55 = 0.91, and the single cells by 50/45 = 1.11. This calculation is done within each sub-cell[23].

To ensure that this process produces realistic output, we ran a simulation, creating a synthetic state-level joint distribution for all 50 states and the District of Columbia, and comparing it to the "true" joint distribution, as seen in the 2014-2018 ACS microdata. The process works well, as shown in Figure A.1 in the Appendix.

[20] The 5-year microdata has 12,962,531 responses among people over the age of 16.

[21] I.e., 45-49 year-old married female citizens with a post-graduate degree in Iowa.

[22] We could also build larger sets of partially pooled models across the country, but did not do so in this paper.

[23] This means that variable order may be important in constructing the synthetic dataset, because we fit the final distribution to the last marginal variable added. In a supplemental analysis, we estimated variable importance for labor force participation and employment and used the following ordering (most important to least): age, education, marital status, race, gender, citizenship. In practice, a different order would probably not make a difference. In the Appendix, we show that our synthetic distribution lines up very closely with a true state-level distribution for all variables, see Figure A.1.

2.1.2 Demographically Adjusted Geographic Predictors

We include geographic predictors in our model, for a number of reasons. We would like to examine geographic variation in our model output, so it is unwise to rely on demographic variation to drive the geographic output alone. Geographic predictors are also a natural way to include industry and occupation data in our model, which is important given the nature of the COVID-19 crisis, affecting different sectors in different ways. Lastly, we want to take advantage of LAUS data in our model estimates.

We include these data points through *demographically adjusted geographic predictors* (DAGPs). Conceptually, it might be easiest to think of DAGPs as ACS or LAUS estimates at the county or census tract level, where each value has been allocated to the different demographic groups inside of each geography. In practice, we build DAGPs going in the opposite direction: building a demographic model for each DAGP, and then constraining the output to add up to a pre-specified target in each geography.

We build a DAGP for each variable listed in the middle column of Figure 2. For industry/occupation variables, the first step is a model using the most recent 5-year ACS microdata, similar to the one described in Equation 1 but with more covariates. As such we generalize the notation. Again, $y_i \in \{0, 1\}$ indicates the response variable for respondent i, and we estimate $P(y_i = 1) = \theta_i$. We use K covariates[24], indexed as $j_1 = \{1, \ldots, J_1\};\ j_2 = \{1, \ldots, J_2\};\ \ldots\ ;\ j_k = \{1, \ldots, J_K\}$; the number of levels for each k factor is indexed using J_k. The association between each of these variables and y can be captured through a series of varying intercepts. The varying intercepts for factor K are denoted $\alpha^k_1 \ldots \alpha^k_{J_k}$, and so the resulting non-nested (crossed) equation is:

$$\theta_i = \text{logit}^{-1}\left(\alpha^0 + \sum_{k=1}^{K} \alpha^k_{j_{k[i]}}\right) \tag{6}$$

$$\alpha^k \sim \text{Normal}(0, \sigma^2_k), \text{ for } j_k = \{1, \ldots, J_K\} \tag{7}$$

Where α^0 is the overall intercept. This model can be interpreted as a hierarchical version of a logistic regression including "base" effects only[25] We would also like to include two-way interactions between all of the factors as well. The two-way interactions could be included in a similar form, but instead we include them in a second-stage LASSO regression, allowing the model to potentially include any interaction, but regularizing the vast majority of them to zero in a computationally efficient manner[26]. The resulting LASSO output[27] is labeled θ^*_i. Applying these models to the joint demographic distributions described earlier moves from the individual respondent θ^*_i to the *cell*-level respondent θ^*_c, indexed on c for C cells in every geography.

The last step is to adjust these to add up to the target number in each geography. We do so through a logistic intercept correction. In any geography g, we treat the most current ACS estimate as the "true" target ξ_g. We derive the adjusted DAGP estimate θ^{**}_c for each cell as follows:

$$\delta_g = \text{argmin}\left(\text{abs}\left(\xi_g - \sum_{c \in C} \text{logit}^{-1}\left(\text{logit}\left(\theta^*_c\right) + \delta\right)\right)\right) \tag{8}$$

$$\theta^{**}_c = \text{logit}^{-1}\left(\text{logit}\left(\theta^*_c\right) + \delta_g\right)\ \forall\ c \in C \tag{9}$$

[24] As described in Figure 2: age, education, race, marital status, citizenship, gender, state, and region. We also include coarsened variables: race (white, non-white) x education (college, non-college); race (white, non-white) x gender, and marital status x gender.

[25] In practice, it is more computationally efficient to include groups with only two levels (i.e., gender, marital status, citizenship) as binary predictors. We do so but leave the cleaner notation here.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Vol. 2. New York: Springer, 2009.

[27] In the future, we

where abs() is the absolute value function and argmin() is a function that finds the δ that minimizes the expression. This process simply applies a constant logistic adjustment δ_g to every cell in a specific geography, to ensure that the total number of people in a specific industry/occupation/employment type is correct[28].

For DAGPs based on the LAUS, we have to go through a couple of additional steps. We are interested in creating DAGPs for labor force participation and the employment rate. The ACS includes these questions, so we go through similar steps, and we correct to the most recent ACS targets. After that is completed, we do an additional round of adjustments, making sure each θ_c^{**} adds up to the correct county-level LAUS number, which can excessively be labeled θ_c^{***}. Lastly, we make sure that the LAUS employment rate DAGP is computed with labor force participation as a denominator, to make it more compatible with the models built in Section 2.2.

2.2 Fit CPS Models

We fit three models using CPS microdata:

1. Labor force participation, among the citizen non-institutionalized population (CNIP) over the age of 16.
2. Employment, among respondents in the labor force.
3. People who are "at work" or absent for usual reasons, among respondents who are employed.

All three can be thought of as probabilities on a 0 to 1 scale that will be applied to cells in the poststratification dataset. As such, they build on top of each other, allowing us to examine various aspects of the labor force. For instance, if we want to examine employment rate using the 16+ CNIP as a denominator, we can multiply the outputs of Models 1 and 2.

Labor force participation and employment are characteristics we clearly want to examine. We build the at-work model for a few reasons. First, the April 2020 jobs report included an important endnote about this topic. Survey interviewers were instructed to classify all employed persons absent from work due to COVID-related business closures as unemployed (on temporary layoff), but that did not consistently happen. CPS administrators decided to keep the data coding consistent, but noted that this would have raised the overall unemployment rate by nearly 5 points. Building both the employment and at-work models allows us to examine employment both ways[29]. Second, this shows an example of using our framework to look deeper than the topline labor force / unemployment rates. Our framework can be used to examine increasingly detailed questions about the labor force, including topics such as part-time work, industry-specific job losses, and others. For these latter questions, we do not have the benefit of the LAUS DAGPs, so the results should be interpreted as having wider uncertainty, but they are still potentially valuable topics, and can be expanded upon in future iterations of this work.

We fit each model sequentially, iteratively adding pieces one set at a time. We could instead build a model that incorporates all of the various pieces at the same time, but we found the following format to fit our needs and be computationally efficient. We are more interested in fitting these models quickly—we plan to fit them for every month of the CPS going back many years—but others may want to build a fully Bayesian model that incorporates everything at once, which would be more appropriate for estimating uncertainty through the multiple steps[30]. The various steps are:

[28] Even the 5-year ACS estimates have fairly low sample size at the census tract level and are quite noisy as a result. As such, we lightly smoothed the targets, ξ_g, giving each census tract a weighted average of the raw value (weight = 2/3) and the size-weighted average of each of its neighbors (weight = 1/3).

[29] Here, we use the coding scheme described in the jobs report endnote, where people who were absent from work for "other reasons" are treated as unemployed.

[30] A full computation of uncertainty in our process would require accounting for the CPS models as well as the other pieces of the process,

A. Baseline geographic model, using previous month's LAUS data (this step is skipped for the at-work model).
B. Industry/occupation model, adding the other DAGPs.
C. Demographic model 1: base data in a hierarchical model.
D. Demographic model 2: two-way interactions in a LASSO.

We follow earlier notation in describing each step. For each model, $y_i \in \{0, 1\}$ indicates the response variable for respondent i, restricted to the appropriate set of respondents, and we are interested in estimating $P(y_i = 1) = \theta_i$ in the various steps. For the labor force participation and employment models, we first fit Step A:

$$\theta_i^A = \alpha^A + \beta^A x_{baseline}, \tag{10}$$

where $x_{baseline}$ is the LAUS-based DAGP for the relevant model. In other words, we first fit a simple linear probability model, estimating the relationship between last month's county-level LAUS DAGP and this month's data[31]. Next we move to Step B, adding the other DAGPs:

$$\theta_i^B = \text{logit}^{-1}\left(\alpha^B + \text{logit}\left(\theta_i^A\right) + \boldsymbol{\beta^B} X_{DAGP}\right) \tag{11}$$

Notice we move to logistic regression for this and the remaining steps, and θ_i^A is included with slope = 1 on the logit scale, so that the relationship between it and the response is linear in the logistic regression. As we progress through the multiple steps here, we regularize model output at every step. If we were to estimate new slopes for the preceding model outputs, our estimates would progressively shrink toward the national average; constraining these slopes to equal 1 prevents that form of over-regularization from happening. $\boldsymbol{\beta}^B$ are a set of slopes for X_{DAGP}, i.e., the set of industry and occupation DAGPs (along with the LAUS DAGP that was not used as $x_{baseline}$ in Step A). This equation is fit through a LASSO regression, regularizing most of $\boldsymbol{\beta}^B$ to zero. In the at-work model, we fit a similar equation, excluding the θ_i^A term, with no more changes in Steps C and D. Step C is a hierarchical logistic regression on demographics, following a similar form and notation to Equation 6:

$$\theta_i^C = \text{logit}^{-1}\left(\alpha^C + \text{logit}\left(\theta_i^B\right) + \sum_{k=1}^{K} \alpha^k_{j_{k[i]}} + \beta x_{claims}\right) \tag{12}$$

$$\alpha^k \sim \text{Normal}(0, \sigma_k^2), \text{ for } j_k = \{1, \ldots, J_K\} \tag{13}$$

Again we pass in the output from the previous step, θ_i^B, with slope = 1 on the logit scale, and we use the same demographics, state, and region variables that were listed earlier[32]. We also include weekly unemployment insurance claims data as a state-level variable, x_{claims}[33]. Lastly, we search for the set of two-way interactions

including creating the synthetic joint distribution, sampling error in the ACS, projecting the ACS to 2020, and so on. Regardless, building a Bayesian model for this piece alone would be fairly straightforward, e.g., in the Stan probabilistic programming language: Stan Development Team. *Stan: A C++ Library for Probability and Sampling, Version 2.23*. 2020. URL: http://mc-stan.org/.

[31] θ_i^A is inserted into the next step on the logit scale, which deals with implied probabilities that are outside the $[0, 1]$ bound. β^A is restricted to equal 1 in the labor force participation model, which fit the local data more efficiently. We also tested using logistic regression in this step. The results were very similar, but a simple linear trend fit the data a bit more closely. Given that DAGPs are imprecise in rural areas due to lack of geographic indicators, we were also worried that extending a logistic trend might have unintended consequences.

[32] We add coarser age groupings—16-24, 25-44, 45-64, 65 or over—as another set of variables here, in case the larger sample size from these groupings allows for stronger relationships to be found, particularly when searching for interactions in Step D.

[33] We use the total number of claims from the four weeks surrounding the 12th of the month, standardized by mean-centering and dividing by two standard deviations, to make the coefficient comparable to those of binary predictors Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press, 2007

in Step D:

$$\theta_i^D = \text{logit}^{-1}\left(\alpha^D + \text{logit}\left(\theta_i^C\right) + \boldsymbol{\beta^D} X_{inter}\right) \tag{14}$$

$\boldsymbol{\beta}^D$ are a set of slopes for X_{inter}, which are indicator variables representing all two-way interactions between demographics, state, and region. This equation is again estimated with a LASSO regression, with most of $\boldsymbol{\beta}^D$ regularized to zero. This completes the fitted model.

Our model, then, accounts for last month's LAUS data, industry and occupation data at the geographic level (adjusted for demographic variation), and demographics, state, and region with two-way interactions between them. It is important to take a moment to discuss variables that are *not* included as covariates in the model: income and self-reported industry / occupation within the CPS microdata itself. These are important covariates, but inside of any single CPS survey the responses mainly come from people *who have jobs*, and thus they can not be used as predictors of labor force participation or employment status. In theory, it may be possible to use a research design that leverages the rolling sample of the CPS to examine panel respondents who answered those questions in a previous survey, but that is quite a different research design than what we propose here. Instead, we include industry / occupation as DAGPs, so the model can pick up these trends indirectly and project them onto the unified poststratification datasets described in Section 2.1.

2.3 Project to Poststratification Data

Recall from Section 2.1 that we have three full poststratification datasets: at the state, county, and census tract level. Each dataset includes the same covariates that were used in fitting the CPS models. We project the labor force estimates to these data iteratively, starting from the smallest geography (tract) and adjusting them so they add up to the best number available at a higher aggregate level.

First, we "score" each step of each model on the census tract poststratification data. In other words, we apply the equations fit in Section 2.2 to the data to come up with estimates for each sub-cell. Note that the DAGPs at the tract level have a wider range of values than those we used in fitting the CPS models, because the latter were only available at the county, metro or state level. This is a feature that we explicitly want to exploit in our estimates, because we are interested in sub-county variation, but it is important to understand we are simply *assuming* that the higher-level DAGP relationships also hold at the tract level.

Next, we use multiple stages of corrections, done on the logistic scale as in Equation 8. We adjust our estimates towards targets that we consider increasingly accurate, the higher up the chain. We use the following sequence:

- County-level targets, which are based on the "scored" county-level poststratification dataset.
- National demographics, one variable at a time and for two-way interactions between variables. These are based on the "scored" state-level poststratification dataset, which is aggregated to produce national numbers. In practice, these are close to the official weighted data published by the CPS, but they account for survey non-response and small-sample national cells.
- If LAUS data is available at the county level (for the labor force and employment models), we adjust toward those numbers for each county.
- If LAUS data is available at the state level, we do the same thing for each state.
- Overall national target, based on the official CPS estimate[34].

[34]We also adjust the overall population to match the national CPS estimate for the 16+ CNIP. For example, the 2018 ACS estimates the

This elaborate set of post-model adjustments may seem excessive, but the motivation is as follows. From our perspective, statisticians at the BLS and LAUS have spent many years developing methods to estimate these statistics at the national and geographic levels, incorporating auxiliary datasets that we do not include in our model. Our method is designed to complement the CPS and LAUS, and we do not have reason to believe that our method should be used to compete with or supplant the official data. As such, these adjustments iteratively correct to the various levels of CPS and LAUS data, progressively moving towards the numbers that we believe are the most accurate. At higher levels, we have numbers that will be very close to the official data, and the purpose of our method is to examine deeper subgroups, i.e., census tracts and demographic groups within state and county.

At the same time, it is important to note that our numbers are not substantially driven by these post-model corrections. For instance, Figure 3 compares our pre-adjustment county-level estimates to the LAUS county-level data. In February and March, which are closer to "normal" months, they are very similar, which builds confidence in our estimates more broadly. Even in April 2020, when the one-month change was larger than at any other time in the history of the series, labor force participation estimates are very similar. The model's output for the unemployment rate differs from LAUS in some places, but we pick up quite a lot of the variance. The "final" model output adjusts to match the LAUS, and we expect "pre-adjustment" estimates to be closer to the LAUS in upcoming months, due to this correction and the (expected) slower pace of monthly change.

3 Example Results

Our method produces a database of labor force estimates, disaggregated by census tract and demographic group, that are consistent with higher-level official sources from the CPS and LAUS. The benefit of the disaggregated data is that it allows for deeper and more flexible analysis of the data, through re-aggregation for any grouping of interest. This section shows examples of what can be seen from this data.

Figure 4 shows census-tract level unemployment rates across the country. We immediately see wide and important geographic variation in unemployment across the country. The national unemployment rate is 14.4%[35], which varies greatly both across and within states. At the state level: Nevada, Michigan, and Hawaii all have unemployment well above 20%. Other parts of the country are much lower, like the upper Great Plains region with 9% unemployment as a whole[36]. Variation within state is often much larger, as can be seen easily in the map and even more so when you zoom into different cities. We show examples in the next few figures, but examining all of the relevant areas is beyond the scope of this paper. If you would like to zoom in on any specific areas, a high-resolution version of this map can be found here.

The geographic differences shown here are important, but close inspection of the map reveals limitations of our approach. The model shows unrealistic state effects in some regions—for example, in Connecticut's low unemployment rate compared to its neighbors, or the stark contrast between eastern Minnesota and western Wisconsin. Remember that our approach matches the county-by-county LAUS data, which shows the same state effects. Our model allocates labor force statistics across each county based on the underlying demographics and tract-level covariates; sometimes this will minimize county- and state-level bordering effect, and other

16+ citizen population to be 262,216,823 people. The April 2020 16+ CNIP estimate is 259,896,348, smaller than the ACS number in part because it reflects the non-institutionalized population. We adjust the ACS data in our poststratification tables to match the CPS number, i.e., by multiplying each cell by 259,896,348 / 262,216,823 = 0.99. A more elaborate model could be built to estimate cell-level adjustments regarding the prevalence of the non-institutionalized population and population change from 2018 to 2020, but we use this simple adjustment here.

[35] Among the labor force, not seasonally adjusted.

[36] Montana, North Dakota, South Dakota, Nebraska, Iowa, and Minnesota.

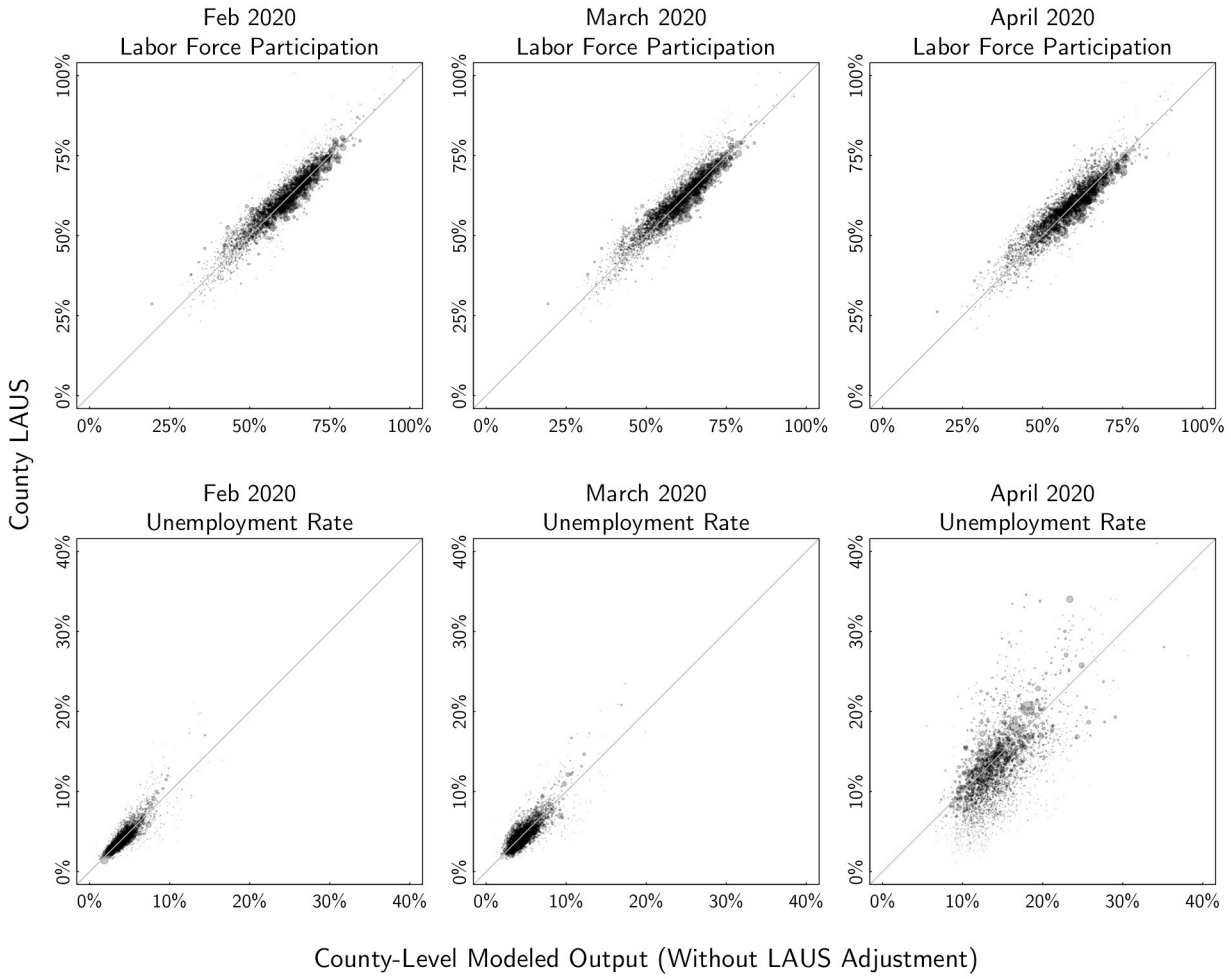

Figure 3: *We compare county-level aggregation of our raw modeled output (x-axis) to the county-level LAUS data (y-axis), released roughly 3 weeks after the CPS microdata. The size of each circle reflects the denominator: county CNIP over the age of 16 (in the labor force participation plots) or the size of the labor force (in the unemployment rate plots). In February and March, which are closer to "normal" months, raw estimates are close to the LAUS. Even in April 2020, when the one-month change was larger than at any other time in the history of the series, labor force participation estimates are very similar. The model's output for the unemployment rate differs from LAUS in some places, but we pick up quite a lot of the variance. The "final" model output adjusts to match the LAUS, and we expect "pre-adjustment" estimates to be closer to the LAUS in upcoming months, due to this correction and the (expected) slower pace of monthly change.*

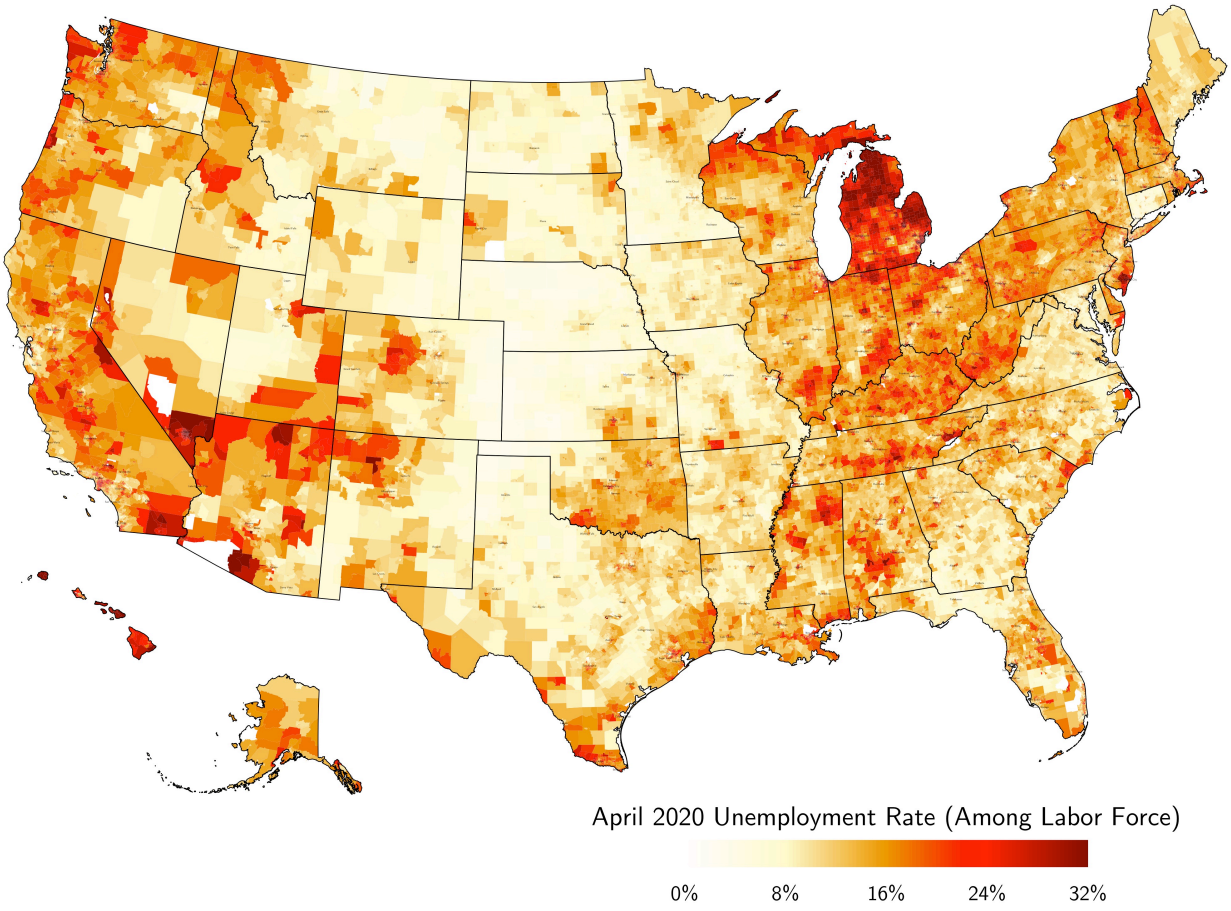

Figure 4: *April 2020 unemployment rate by census tract. The national (not seasonally adjusted) rate is 14.4%, but geographic variation is high, both across and especially within states. The next set of figures zoom in on example cities, and a high-resolution version of this map can be found here. Close inspection of the map reveals limitations of the model, namely unrealistic state-level discontinuities. The impact of these discontinuities are visually over-emphasized, however, because they appear in geographically large rural areas with a small number of labor force participants.*

times it will not. With that said, many of the border areas are rural, where the map gives a lot of visual weight to the large geographic area even though the number of jobs is small.

We can see important within-state variation more clearly by zooming in to different areas. Figure 5 looks at New York City and the surrounding area, relevant because it was the region with the highest number of cases of COVID-19[37]. Instead of looking at the raw unemployment number, we look at the number of jobs lost between February and April, choosing February as the baseline because the survey was conducted before the large-scale economic impact of COVID-19 was felt. We think this is a more informative metric; while unemployment was going up, labor force participation dropped from 63.3% to 60.0% in this time period, its lowest point since 1973.

The map on the top-left shows the five counties that comprise New York City; this level of granularity is not sufficient to understand what was happening in the city, as all five counties uniformly lost about 19% of their

[37] These maps were built using the ggmap package in R: David Kahle and Hadley Wickham. "ggmap: Spatial Visualization with ggplot2". In: *The R Journal* 5.1 (2013), pp. 144–161. URL: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

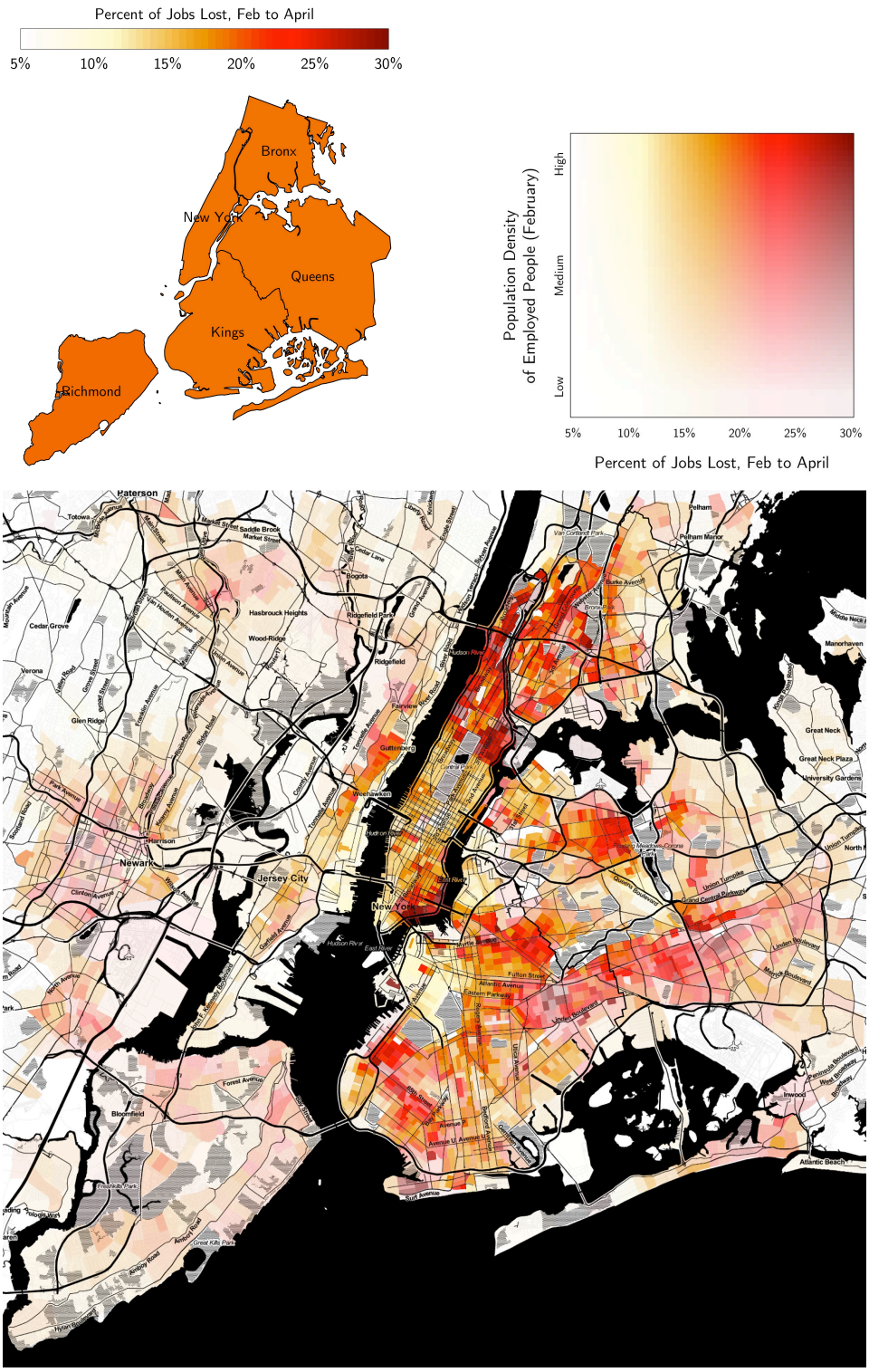

Figure 5: *Percent of jobs lost between February and April 2020 in the New York City area. County-level estimates are not sufficient to understand what was happening in the city, as all five counties uniformly lost about 19% of their jobs. Our model suggests that parts of the Bronx, Brooklyn, and Queens were hit hard economically, with as many as 35-40% of people in many neighborhoods losing their jobs. Please note that these data represent households where people live, not business locations.*

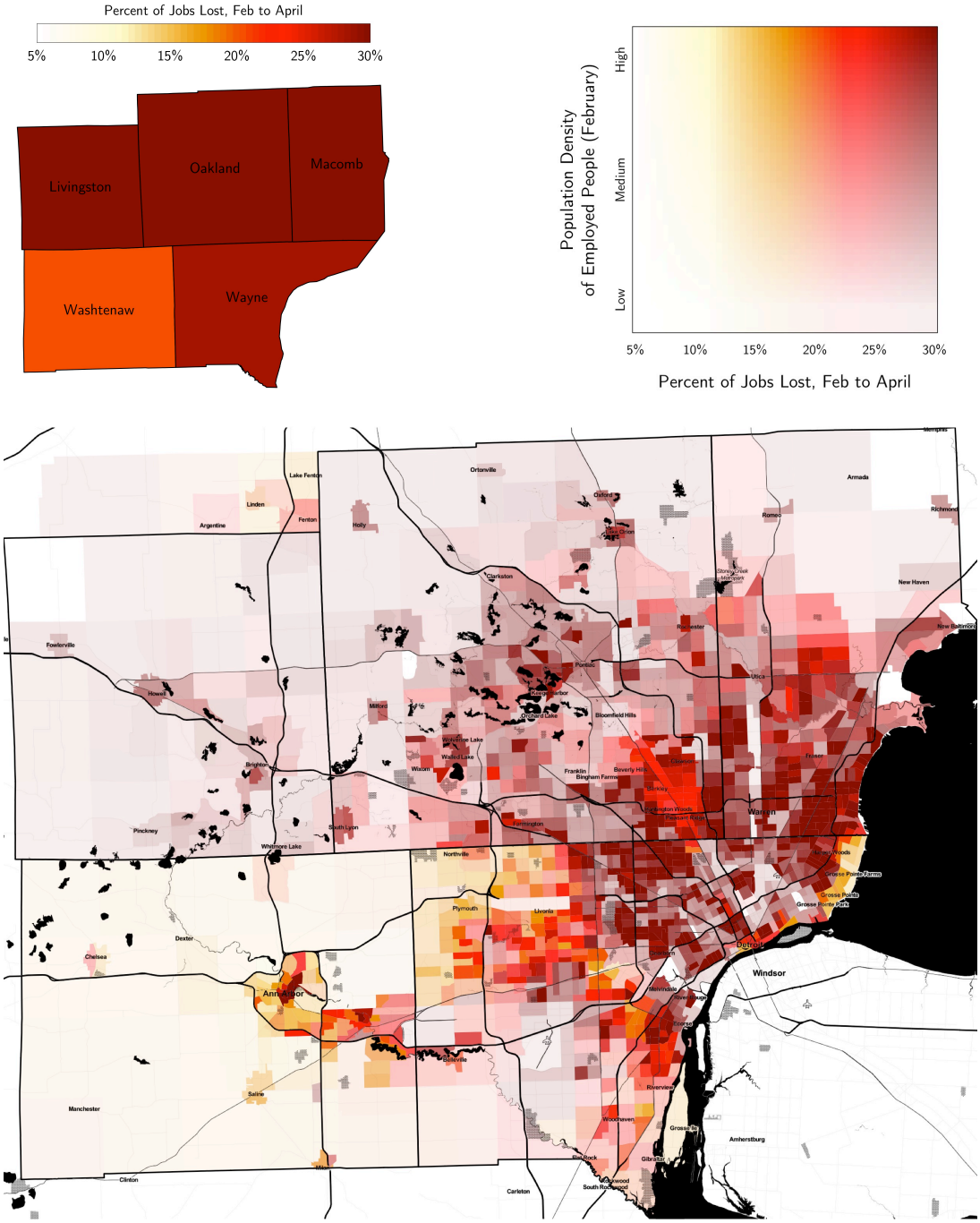

Figure 6: *Detroit is one of the hardest hit cities in the country, with parts of the city reaching well over 50% of jobs lost. Again, the county data is useful but incomplete, as it averages over diverse terrain, often mixing inner cities and educated suburbs that are in very different situations. Both this figure and the previous one consider people employed whether they were "at work" or not; if we restrict to people who were at work, the damage is even greater. Like the previous figure, these data represent households where people live, not business locations.*

jobs. The larger map on the bottom shows how this masks a great deal of within-county variation, with the color scheme now making low-density places more transparent, placing more emphasis on areas with more jobs. Our model suggests that parts of the Bronx, Brooklyn, and Queens were hit hard economically, with as many as 35-40% of people in many neighborhoods losing their jobs[38].

Figure 6 shows the same types of maps around Detroit. This is one of the hardest hit cities in the country, with parts of the city reaching well over 50% of jobs lost. Again, the county data is useful but incomplete, as it averages over diverse terrain, often mixing inner cities and educated suburbs that are in very different situations. Even though these maps show incredibly high numbers of lost jobs, they actually *underestimate* the damage because we use the official definition of employment to make the maps comparable to the LAUS county data. If we used the "at work" definition, the numbers would be substantially higher. This is easy to do in our framework, and we move to that definition in the next set of graphs.

Before doing so, it is important to step back and remember what this data reflects. The lowest level of "official" geographic data used in our models were county-level LAUS estimates[39]. All of the variation in our model below the county level is driven by our model: it sees that certain demographic groups (communities of color, people without a college degree) are losing their jobs and projects those inferences inside of each county. The model also uses historic tract-level employment, industry, and occupation data. We think this reveals important differences at the sub-county level, but the model is only an estimate, and there will certainly be places where it is wrong. Because of this, we describe the data as suggestive instead of definitive, particularly now when labor force statistics are moving at such a rapid pace due to unprecedented shocks.

Next, we move on to demographics. While national demographic analysis is possible using standard CPS data, there is usually not enough sample size to produce reliable demographic estimates at the state level. Our model facilitates this analysis, as shown in Figure 7. Our background is in political analysis, and with Michigan being one of the most important swing states in the 2020 election as well as one of the hardest hit states economically, we compare different groups' levels of job losses to their voting patterns from the 2016 election. The x-axis shows how much support Hillary Clinton got from each group, according to data provided by Catalist, a private firm[40]. The y-axis focuses on *total* jobs lost, now treating people who were employed and "at work" as having jobs, as suggested in the April 2020 CPS jobs report[41]. And the size of the gray bubble indicates the size of that group in the population. The left-hand plot examines one demographic group at a time, limited to groups that are greater than 3% of the population. Broadly speaking, groups that supported Clinton bore much of the brunt of the recent job losses. African Americans, Hispanics, young people, unmarried people, and people living in cities constituted much of Clinton's support, and those groups range from 37 to 43% of their February jobs being lost. In comparison, "only" 20% of college-educated Michiganders lost their jobs, while the most Trump-leaning group, people in rural areas of the state, lost 30%. The right-hand plot shows interactions between these variables (all of the gray bubbles), highlighting only a few of them to keep the text readable. Looking at the interactions provides more detail and clarifies the situation, e.g., that black non-college women lost 47% of their jobs, while white college men lost only 18%. We can also now see some swing or Trump-leaning groups with massive job losses, like white women over the age of 65, or white

[38] Because we are using the CPS household survey, this reflects where survey respondents live, not business locations.

[39] The LAUS publishes some data at sub-county levels, i.e., estimates for cities with population greater than 25,000. That data is not produced for all cities, though, and does not show geographic variation within cities. Still, it is valuable information, and future iterations of our model could incorporate that data to improve estimates.

[40] One of the authors is the Chief Scientist at Catalist, and produced these estimates using a conceptually similar approach to the one described in this paper. See Yair Ghitza. *What Happened Next Tuesday: A New Way To Understand Election Results*. 2019. URL: https://medium.com/@yghitza_48326/what-happened-next-tuesday-e4e6637a4b81

[41] This was done for both the February and April numbers, and we focus on citizens over the age of 18 to make the results more comparable to the voting numbers. These groups are still not entirely comparable, because the voting numbers reflect 2016 voters.

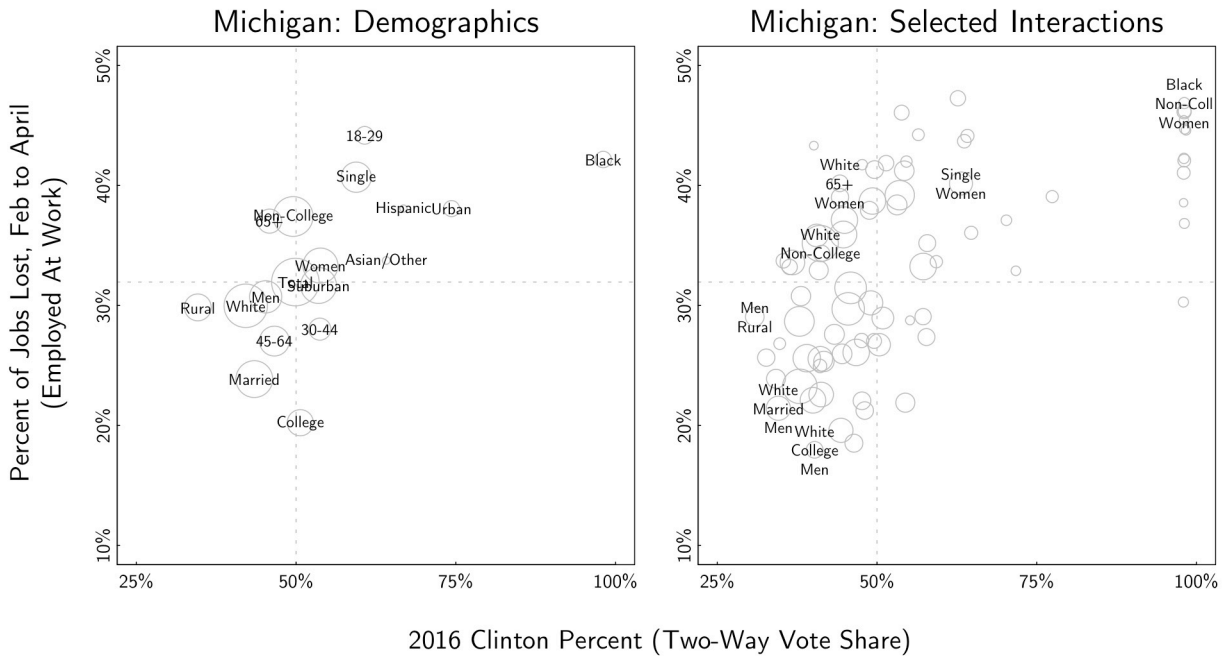

Figure 7: *Demographic analysis in Michigan, one of the hardest hit states in the country and an important swing state in the upcoming 2020 election, showing groups that are 3% of the state or larger. Standard CPS data does not have enough sample size for reliable inference here. The left-hand plot examines one demographic group at a time: groups that supported Clinton bore much of the brunt of the recent job losses. The right-hand plot shows interactions between these variables (all of the gray bubbles), highlighting only a few of them to keep the text readable. Looking at the interactions provides more detail and clarifies the situation, e.g., that some swing or Trump-leaning groups also saw massive job losses. Notice that the axes do not start at zero, to emphasize the relevant variation.*

non-college people writ large. It is too early to tell what impact these job losses will have on the upcoming election, but there is little doubt that a wide swath of people in Michigan have been deeply impacted by these shocks to the labor force.

4 Discussion

Methodological innovations and increased computing power have allowed researchers to incorporate external multilevel data about the populations surveyed and provide reliable modeled insights about subgroups, especially local areas. Our work is motivated by this work and is, as noted in the text repeatedly, designed for speed and broad reasonableness. We are aware of many potential substantive concerns and have many ideas for additional research; an initial discussion of these begin below.

But there is a first concern and caveat that we wish to address directly. We are not labor market economists, and we are working in an area where expertise is transparently important. Methodologically, we have relied as closely as we can on the official statistics. We have tried to keep the modeling directly dependent upon and predictive of standard variables (even in these non standard times). We have made our work transparent. In the normal course of events, we would spend months work-shopping these ideas and reviewing them with colleagues and through networks, but these are not normal times. Nonetheless, we are aware that statistical and data expertise without substantive knowledge is dangerous, and we are seeking criticism from readers of this working paper, to keep this work from falling prey to those problems. But just as we seek focused criticism, we also ask that those who are unfamiliar with these techniques to consider the contribution they can make to localized estimates of the current labor market.

When it comes to our modeled estimates, a critical question is, to use Les McCann and Eddie Harris's phrase, "Compared to What?" We understand that as time moves on, additional sources of detailed local labor force data will become available, current estimates will be revised, and we will collectively have an understanding of the local economic impacts of COVID-19 that is closer to ground truth. Our rough and ready local area estimates can be thought of as a rough preview. Our motivation is adding another option for decision makers and analysts in the trade-off between timeliness and precision. It is our experience in policy and politics that when data analysts refuse to make a needed estimate, policy makers and managers often use crude and obvious short cuts. The survey can't tell you about Hispanic employment in Arizona and Nevada, so let's assume they are about as much above average as Hispanic employment everywhere else. It is our hope that these estimates will help decision makers do better than that as a first cut.

We can imagine these estimates being of interest in a wide range of circumstances. Even before COVID, place-based policies have been the focus of increasing attention. Direct public investment, tax expenditures and grants for businesses and individuals, and regulatory relief vary spatially[42]. It is our particular hope that these estimates will be helpful to federal, state, and local governments as they consider policies to deal with the current economic and health crises.

How do our estimates compare to those developed using alternative approaches, such as the large scale payroll analyses? Can these models be expanded to include more granular historical data explicitly rather than implicitly through the CPS and LAUS data as in the current model? Are there specific areas where these assumptions are not appropriate, and different methods are better? Can these approaches address non-response issues in the CPS that seem increasingly important? Are there specific policy areas where these estimates are of

[42] Benjamin A Austin, Edward L Glaeser, and Lawrence H Summers. *Jobs for the Heartland: Place-Based Policies in 21st Century America*. Tech. rep. National Bureau of Economic Research, 2018.

particular help? What are the formal statistical properties of our estimators? We see many research ideas and potential applications we and others may wish to pursue.

This first version of the working paper only includes models based on the February and April 2020 CPS. We will be running these models again as new data is published. If there is interest, we will also run the models for past months and years to facilitate longer term historical analysis of the labor force.

A Appendix: Supplemental Materials

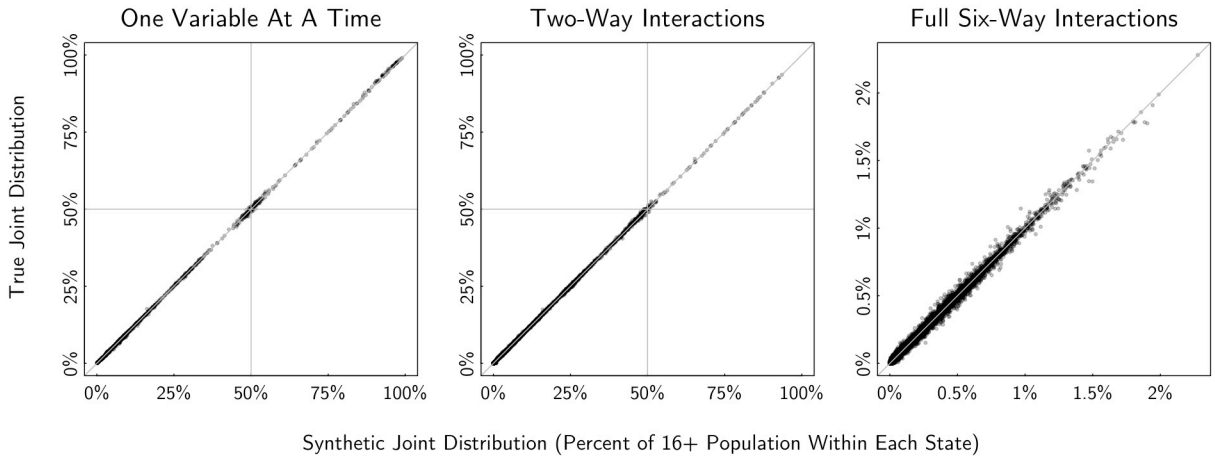

Figure A.1: *Simulation results for Section 2.1.1, where we create synthetic joint demographic distributions for different geographies. Here we ran the process for all 50 states and the District of Columbia, comparing to the "true" joint distribution of each state, as seen in the 2014-2018 ACS microdata. In each graph, the x- and y-axes show the percent of any group inside a single state, first showing variables one at a time (left), then two-way interactions (middle), and finally the full six-way cells (right).*

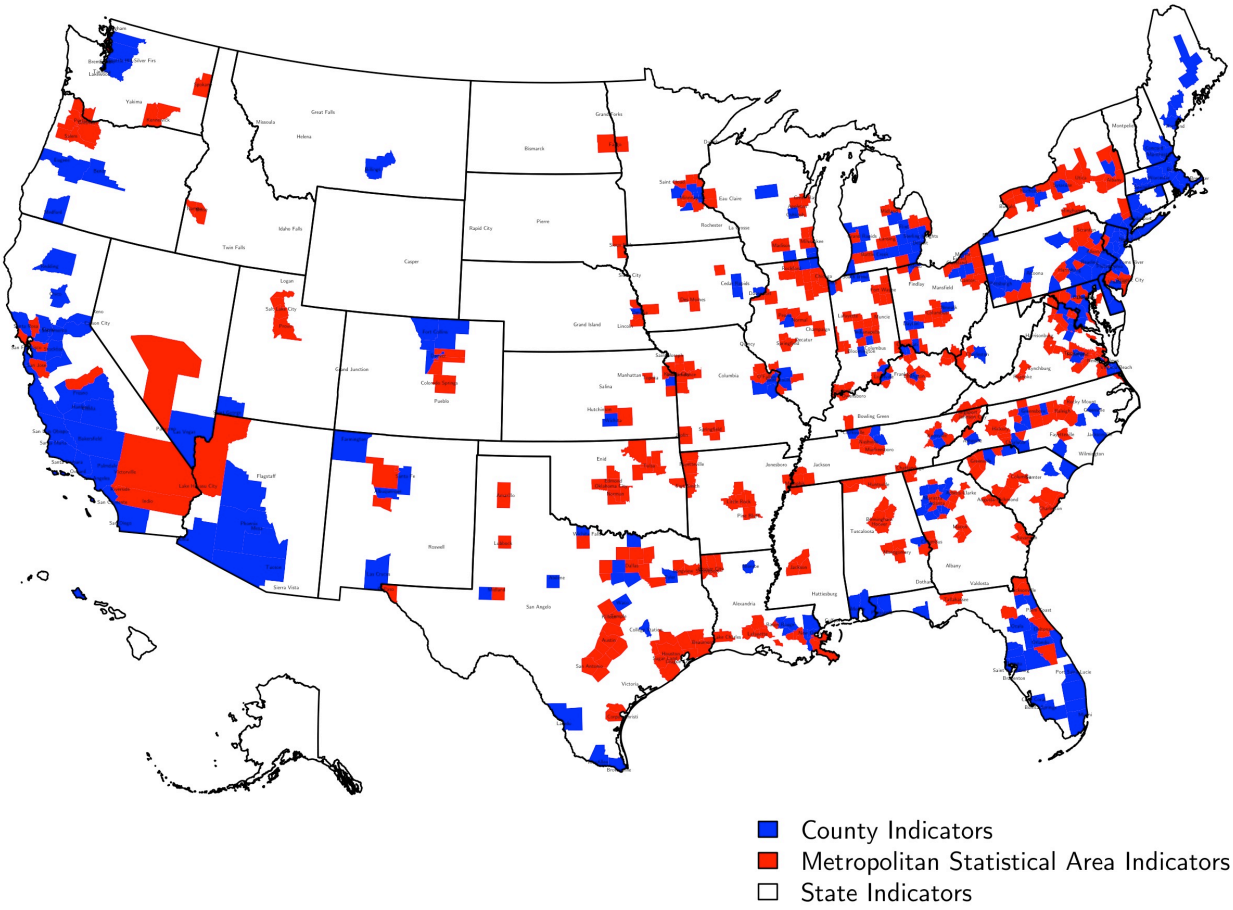

Figure A.2: *Geographic indicators available in the April 2020 CPS microdata. Indicators for county and metro area are included when there are enough responses to avoid privacy concerns. When appending Demographically Adjusted Geographic Predictors (DAGPs) in the CPS data, we use county data where indicators are available (blue counties in the map, 41% of respondents). For respondents that have metro area but not county (shown in red, 34%), we use DAGP data from that metro area, computing values using counties that are not reported in the CPS. For everyone else (shown in white, 25%), we use state-level DAGPs, removing the counties and metro areas from earlier steps. Even in the state-level areas, DAGPs account for state, demographic, and county-exclusion variation in the historical data, as estimated using a model on 5-year ACS microdata.*

References

Austin, Benjamin A, Edward L Glaeser, and Lawrence H Summers. *Jobs for the Heartland: Place-Based Policies in 21st Century America*. Tech. rep. National Bureau of Economic Research, 2018.

Bartik, Alexander W et al. "Labor Market Impacts of COVID-19 on Hourly Workers in Small-And Medium-Sized Businesses: Four Facts From Homebase Data". In: (2020).

Bick, Alexander and Adam Blandin. "Real Time Labor Market Estimates During the 2020 Coronavirus Outbreak". In: *Unpublished Manuscript, Arizona State University* (2020).

Bregger, John E. "The Current Population Survey: A Historical Perspecitve and BLS Role". In: *Monthly Lab. Rev.* 107 (1984), p. 8.

Cajner, Tomaz et al. "The US Labor Market During the Beginning of the Pandemic Recession". In: (2020).

Caughey, Devin and Christopher Warshaw. *Public Opinion in Subnational Politics*. 2019.

Chetty, Raj et al. "Real-Time Economics: A New Platform to Track the Impacts of COVID-19 on People, Businesses, and Communities Using Private Sector Data". In: (2020).

Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber. "Labor Markets During the Covid-19 Crisis: A Preliminary View". In: (2020).

Dunn, Megan, Steven E Haugen, and Janie-Lynn Kang. "The Current Population Survey: Tracking Unemployment in the United States for Over 75 Years". In: *Monthly Labor Review* (2018), pp. 1–23.

Flood, Sarah et al. *Integrated Public Use Microdata Series, Current Population Survey: Version 7.0*. Minneapolis, MN, 2020.

Gelman, Andrew and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press, 2007.

Ghitza, Yair. *What Happened Next Tuesday: A New Way To Understand Election Results*. 2019. URL: https://medium.com/@yghitza_48326/what-happened-next-tuesday-e4e6637a4b81.

Ghitza, Yair and Andrew Gelman. "Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups". In: *American Journal of Political Science* 57.3 (2013), pp. 762–776.

— "Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research". In: *Political Analysis* (2020), 125.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Vol. 2. New York: Springer, 2009.

Kahle, David and Hadley Wickham. "ggmap: Spatial Visualization with ggplot2". In: *The R Journal* 5.1 (2013), pp. 144–161. URL: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kahn, Lisa B, Fabian Lange, and David G Wiczer. "Labor Demand in the Time of COVID-19: Evidence From Vacancy Postings and UI Claims". In: (2020).

Kromer, Braedyn K. and David J. Howard. "Comparison of ACS and CPS Data on Employment Status". In: *Census Working Papers* SEHSD-WP2011-31 (2011).

Kurmann, Andre, Etienne Lale, and Lien Ta. "The Impact of COVID-19 on US Employment and Hours: Real-Time Estimates With Homebase Data". In: *Unpublished Manuscript* (2020).

Lax, Jeffrey R and Justin H Phillips. "How Should We Estimate Sub-National Opinion Using MRP? Preliminary Findings and Recommendations". In: *annual Meeting of the Midwest Political Science Association, Chicago*. 2013.

Leemann, Lucas and Fabio Wasserfallen. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation". In: *American Journal of Political Science* 61.4 (2017), pp. 1003–1022.

Ruggles, Steven et al. *IPUMS USA: Version 10.0*. Minneapolis, MN, 2020.

Stan Development Team. *Stan: A C++ Library for Probability and Sampling, Version 2.23*. 2020. URL: http://mc-stan.org/.

Tedeschi, Ernie and Quoctrung Bui. "America's Employment Losses Might Be Slowing: Job Tracker". In: *The New York Times* (2020).

Zhang, Xingyou et al. "Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System". In: *American Journal of Epidemiology* 179.8 (2014), pp. 1025–1033.