## Title

Environmental Inequalities and Social Protection: Essays in Environmental and Public Economics



**Author**
Weill, Joakim Armand

**Publication Date**
2022

Peer reviewed|Thesis/dissertation

Environmental Inequalities and Social Protection:
Essays in Environmental and Public Economics

By

JOAKIM ARMAND WEILL

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Agricultural and Resource Economics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Michael R. Springborn, Co-Chair

Katrina K. Jessoe, Co-Chair

Marianne P. Bitler

Ashish Shenoy

Committee in Charge

2022

# Contents

# Dissertation Abstract

Environmental degradation, and climate change in particular, are increasingly recognized to be among humanity's largest threats. Calls to protect our "Common Home" are becoming more frequent and resolute. While firmly rooted in science, these declarations often build on the implicit assumption that we are all in the same boat. We are not. Environmental degradation has profoundly unequal impacts on individuals depending on where they were born, how much money they inherited, or the color of their skin. Some people will drown, and others are buying 400-feet yachts.

The dangers posed by environmental degradation and the assurance of unequal impacts present colossal challenges for public policy. Even if mitigation actions can still reduce the worst impacts, scientists warn us that massive damages are already inevitable. Yet, because these damages will not be borne by everyone, enacting effective and fair public policies needs to account for variation in these impacts as well as in people's capacities to adapt to them. How should societies best protect their people against environmental risks?

Figuring out the set of policies to best protect individuals presents a unique opportunity for research, and economics in particular. For all the powerful tools that economics offer, tackling this question requires two substantial departures from the current practice of neoclassical economics. First, recognizing that the concept of (potential) Pareto efficiency as a theoretical guide can only take us so far. Policies always create losers; given that the people who lose from a policy are seldom compensated for their troubles, it is not nearly enough to know that a policy could *potentially* improve the welfare of society *if* the right transfers were implemented. Pareto efficiency is not a morally neutral concept – it always gives more weight to a subset of individuals, even if we often do not know who these individuals are. More complete analyses require accounting for the distribution of costs and benefits within the population in order for moral ideals to be an integral and explicit part of policy discussions.

The second departure is the empirical counterpart of the first one. Most work in applied economics and policy analysis focuses on the estimation of *average* effects, where the average is either taken over a large group of individuals exposed to the policy, over a small group of individuals around some cutoffs (with regression discontinuities), or over a sample of "compliers" that we cannot identify (for instance with instrumental variables). Yet, the very moment we start thinking about who gains and who loses from a policy, average effects become quite inadequate. It is not enough to know that the average effect of a policy is positive. For instance, subsidies for individual solar panels could lead to more installations, but these installations might be concentrated at the very top of the income distribution, which could leave disadvantaged households to pay higher energy prices (due to the rising share of fixed costs in the distribution system, as this is the case in California). Taking compensation and distributional concerns seriously requires the use of less common

empirical tools.

I am fortunate to have started my PhD at a time when economics is undergoing a major shift in both directions. The explicit focus on distributional concerns is becoming more legitimate within the neoclassical framework, and this trend is reflected in environmental economics. While sociologists and epidemiologists have studied environmental justice for over 40 years, environmental economists are only starting to routinely take these distributional concerns into account. Similarly, an explosion of recent work in econometrics is pushing the frontiers regarding the estimation of heterogeneous treatment effects, the development of quantile estimators, and the identification of individual-treatment effects to better understand the distribution of policy impacts within a population. This dissertation builds on both trends to investigate environmental inequalities and social insurance in the United States.

The first chapter investigates the public provision of climate risk information and its distributional impacts on the demand for residential flood insurance. Flooding is among the costliest disasters in the United States, but the demand for federally-subsidized insurance is extremely low. I find that inaccurate flood maps explain a substantial share of low insurance take-up. Using a novel approach based on unit-specific synthetic controls, I estimate that official map updates over the past two decades caused substantial declines in the demand for flood insurance, primarily in neighborhoods with a higher share of African Americans. I quantify the welfare costs of these inaccurate map updates, and estimate the distribution of gains that would accrue from the provision of better flood risk information. The second chapter highlights stark mobility inequalities between income groups during the early stages of the COVID-19 pandemic, at a time when media coverage routinely used *guilt* to convince people to stay home. Guilt may work, but if sustained mobility reflects labor constraints rather than lifestyle preferences, money or a functioning safety net would probably have been better to keep people home. The third chapter emphasizes major shortcomings with the methodology used to estimate the impacts of public policies during the pandemic, and shows that standard robustness tests associated with fixed effects models should be viewed with considerable caution. Finally, the fourth chapter presents early findings regarding the impacts of housing choice vouchers on the residential outcomes and environmental exposure of tenants.

## Perilous Flood Risk Assessments

Flooding is among the costliest natural disasters in the United States. While flood insurance is provided by the federal government and largely subsidized, demand for insurance remains extremely low. This paper assembles the most comprehensive set of files ever compiled on flood risk and insurance to investigate the impacts of information provision on the residential demand for flood insurance. I find that households are sensitive to changes in the floodplain boundaries depicted on the official flood maps, but that map updates over the past fifteen years were largely incorrect, thus providing inaccurate risk and price signals to more than 7 million households. Map updates caused demand for insurance to fall on aggregate, with the largest declines concentrated in neighborhoods with a higher share of African Americans – a pattern largely driven by rezoning these households outside of the high-risk floodplains. In contrast, digitizing previously-available information does not impact insurance take-up, suggesting that the private benefits of improving access to flood risk information are small. Leveraging independent flood risk estimates in a structural model of insurance demand reveals that official map updates decreased welfare by at least $300 million dollars, spread across the income distribution. Correcting flood maps would yield welfare gains exceeding 20 billion dollars annually, primarily in wealthy neighborhoods.

### Income and Mobility Inequalities at the Onset of the COVID-19 Pandemic

In the absence of a vaccine, social distancing measures are one of the primary tools to reduce the transmission of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, which causes coronavirus disease 2019 (COVID-19). We show that social distancing following US state-level emergency declarations substantially varies by income. Using mobility measures derived from mobile device location pings, we find that wealthier areas decreased mobility significantly more than poorer areas, and this general pattern holds across income quantiles, data sources, and mobility measures. Using an event study design focusing on behavior subsequent to state emergency orders, we document a reversal in the ordering of social distancing by income: Wealthy areas went from most mobile before the pandemic to least mobile, while, for multiple measures, the poorest areas went from least mobile to most. Previous research has shown that lower income communities have higher levels of preexisting health conditions and lower access to healthcare. Combining this with our core finding—that lower income communities exhibit less social distancing—suggests a double burden of the COVID-19 pandemic with stark distributional implications.

### Researcher's Degrees-of-Freedom: Evidence From the Pandemic Policy Evaluations

The COVID-19 pandemic brought unprecedented policy responses and a large literature evaluating their impacts. This paper re-examines this literature and investigates the role of researchers' degrees-of-flexibility on the estimated effects of mobility-reducing policies on social-distancing behavior. We find that two-way fixed effects estimates are not robust to minor changes in usually-unexplored dimensions of the degree-of-flexibility space. While standard robustness tests based on the sequential addition of covariates are very stable, small changes in the outcome variable and its transformation lead to large and sometimes contradictory changes in the estimates, where the same policy can be found to significantly increase or decrease mobility. Yet, due to the large number of degrees-of-flexibility, one can focus on a set of results that appears stable, while ignoring problematic ones. We show that recently developed heterogeneity-robust difference-in-differences estimators only partially mitigate these issues, and discuss how a strategy of identifying the point at which a sequence of ever more-stringent robustness tests eventually fail could increase the credibility of policy evaluations.

### Housing Assistance, Neighborhood Quality, and Environmental Inequalities

This paper estimates the dynamic impacts of federal housing assistance on the residential outcomes of program recipients. Combining restricted administrative data on all federal housing choice voucher users followed over the past two decades with a novel dataset of environmental and economic measures at the neighborhood level, we first uncover evidence of dynamic selection into treatment: prior to receiving a voucher, future voucher users move to increasingly poorer neighborhoods. We then find that the effects of vouchers on residential outcomes also exhibit sharp dynamics: recipients initially use vouchers to move to poorer neighborhoods before moving to progressively wealthier areas, with near-constant increases in average neighborhood income of about 1.25% per year. All race and ethnic groups experience modest but significant improvements in air quality, while White voucher recipients benefit from lower exposure to toxic release facilities. Overall, our results highlight that housing vouchers successfully steer tenants towards higher quality neighborhoods, but that these benefits take time to materialize.

# Acknowledgements


> Il n'existe aucun moyen de vérifier quelle décision est la bonne car il n'existe aucune comparaison. Tout est vécu tout de suite pour la première fois et sans préparation.
>
> ---
>
> L'insoutenable légèreté de l'être
>
> Milan Kundera

This PhD has been a formidable experience. Although an experience that lasts for several years is not so much an experience, and more of a slice of life. A nineteen percent slice, in my case. I did not know how much I would enjoy this slice going in. Kundera captured our inability to assess counterfactuals, but this slice was over and above what I could have imagined – *ex post*, it was the perfect choice. I have many people to thank for making it so.

Mike Springborn was always available to discuss my – still nebulous – ideas. He provided me with unwavering guidance and support, and encouraged me to pursue the questions I was most excited about. Katrina Jessoe helped me transform ideas into research. She demonstrated how scientific integrity takes precedence over bureaucratic concerns, and guided me through the job market process. Marianne Bitler shared some of the most insightful advice that an empirical researcher could hope to hear. She took the time to understand my data, and then explained it to me. Ashish Shenoy helped me turn my research into economics, and enabled me to discard a fair number of infeasible projects to focus on the viable ones. Pierre Mérel had the precision and technical clarity that my laziness sometimes attempted to bypass. You all showed me that economists can be both brilliant and kind.

Beyond the professors most closely involved with my dissertation, the faculty at UC Davis and in the ARE department in particular has been nothing short of incredible. I owe a lot to the countless conversations I had with Mark Agerton, Tim Beatty, Steve Boucher, Bulat Gafarov, Dalia Ghanem, Travis Lybbert, Kevin Novan, Matt Reimer, Jim Sanchirico, Aaron Smith, and Jim Wilen. I thank Rachael Goodhue for telling me about the ARE graduate program and inviting me to apply during her vacation in Paris, and Rich Sexton for hosting amazing ski trips in Tahoe. I also benefited from conversations with Colin Cameron, Takuya Ura, Monica Singhal and Jenna Stearns in the Economics department, Cristina Moya and Monique Borgerhoff-Mulder in the Anthropology department, and Noli Brazil in the Human Ecology department.

Being part of the ARE program was a unique opportunity to meet amazing people I am fortunate to call my friends. I treasure the many coffees and lunch breaks with Miki Doan – I'm glad you eventually chose Davis, and looking forward to trying better restaurants once we can afford it. I had incredible conversations over wine with Caitlin Kieran, Eric Smyth, Ed Whitney, Charlotte Ambrozek, Ethan Krohn, Jess Rudder, Derek Rury, Laura Meinzen-Dick, and Maxime Depalle (you completed your dissertation way too fast). I still think that Californian reds are arrogantly overpriced and sweet. Ben Dawson, François Castonguay and Adam Soliman got me into IPAs and made the first year unforgettable. I hope to see more of you on the East coast. Thanks to Julian Arteaga and Curtis Morill for being great buddies, Danea Horn for establishing a cool restaurant on the side (Davis needed it), Karen Ortiz Becerra and Xiurou Wu for making our commitment device group, and Ting Bai, Pierce Donovan, Zheng Fu, Rachel Jones, Tina Kotsakou, Emily McGlynn, Anderson Ospino, Adelaida Ortega, Zach Rutledge, Sarah Smith, Scott Sommerville, and Bret Stevens.

Big thanks to Matthieu Stigler for teaching me a ton of coding skills (and bringing some sarcastic humor to this side of the Atlantic), and to Madeline Turland and Reid Taylor for being really fun to work with. I also learned a lot from working with Ellen Bruno, Marshall Burke, Olivier Deschenes, Frank Loge, and Amanda Rupiper. Thanks to John Voorheis for sharing his knowledge of administrative data, being so open to new ideas and helping me navigate complex governmental processes, and to Briana Ballis, Lionel Cosnard, Penny Liao, Matt MacLachlan and Emmanuel Paroissien for insightful discussions and comments on early versions of my work.

The analyses underlying this PhD could not have been performed without the immense support of Arnon Erba, Jeff Goettsch and Laurie Warren. Your help and expertise has been invaluable.

My work benefited from financial support from the John Muir Institute of the Environment, the Center for Data Science and Artificial Intelligence, the UC Davis Provost Dissertation Fellowship, and the ARE Department first year fellowship.

I thank my family for not asking how much longer it would take me to write this dissertation – and for listening to my intermittent complaining without (vocal) judgments. Mom, you are a model for what scientists should look like – curious, passionate, and uncompromising with the occasional jerks along the way. Dad and Arlette, your social work is an inspiration and continues to help me put things in perspective. Ben, thank you for your gentle but effective support during this PhD – you made me avoid some serious career *faux pas*. Justine, Youssef, Gillou, Baptiste, thank you for unforgettable adventures in California and beyond. Maureen, you made this PhD feel like a side quest – the real achievement was the fun we had around it. I love you all. Thank you.

# Chapter 1

# Perilous Flood Risk Assessments

## 1.1 Introduction

Communities, firms and individuals face environmental risks that are both evolving and hard to perceive. Most notably, climate change is increasing the risks of natural disasters. The costs of these risks are enormous: in the last decade alone, natural disasters in the United States caused more than 800 billion dollars in economic damages, and these impacts will further increase due to rising sea-levels, extreme local precipitation, and abnormally dry conditions (Westra et al., 2014; Hinkel et al., 2014; Union of Concerned Scientists, 2017). Insurance products and other adaptation strategies are often available to mitigate these risks, but multiple economic barriers prevent investments in climate resilience (Kousky, 2014; Hallegatte, 2014; Chambwera et al., 2014; The United Nations Environment Programme, 2021; Kahn, 2021). In particular, the evolving information surrounding climate risk and information acquisition costs can substantially limit the scope of adaptation (Moser and Ekstrom, 2010; Eisenack et al., 2014). Even when information about climate risk is accessible, inequalities in adaptive capacities raise distributional concerns (Shi et al., 2016; Byskov et al., 2021).

While the U.S. federal government does not systematically offer protection against environmental risks, one prominent exception exists: flooding. Flooding is among the costliest natural disasters, leading to substantial infrastructure damages, health costs, and reductions in consumption and well-being (NOAA, 2020; Watkins, 2012; Allaire, 2018; Ouazad and Kahn, 2021). In order to increase resilience to flooding, the federal government provides flood insurance to homeowners and renters through the National Flood Insurance Program (NFIP), which underwrites around 95% of all active insurance policies (Kousky et al., 2018). The functioning of the NFIP crucially relies on the provision of flood maps, which depict the spatial extent of the 100-year floodplain, or "high risk zone".[1] Initially paper-based, most flood maps were digitized since 2005 and currently cover close to 90% of U.S. households. Updating these maps involves continued investments that make the flood risk mapping program one of the most expensive cartographic efforts in recent history.

Yet, despite the broad coverage of flood maps and the wide availability of subsidized insurance policies, the demand for flood insurance has remained extremely low: less than 5.5 million residential properties were covered by flood insurance in 2018, when previous work found that at least three times as many households face substantial flood risk (Wing et al., 2018). Such disconnect raises the question of whether individuals know what their exposure to flood risk actually is, and whether the substantial federal investments in flood risk information dissemination are working (Kousky, 2016; Office of Inspector General, 2017).

I consolidate the most comprehensive set of files ever assembled on flood risk and insurance demand to study how the public provision of flood risk information evolved over the past two decades; how this information impacted private investments in residential flood insurance; and how official flood risk assessments differ from the best available science. Obtaining and compiling these data involved four separate Freedom of Information Act requests, multiple meetings with current and former government officials to identify and access the location of cartographic records, and extensive data processing to ensure the comparability of records obtained from different sources. My final dataset comprises all flood maps updated by FEMA over a 15-year roll-out, linked to geolocalized data on the entire residential housing stock in the contiguous U.S., with administrative records from NFIP on all flood insurance policies and independent flood risk estimates from a state-of-the-art model. Importantly, the data allow me to observe both the reclassifications of properties inside and outside of the FEMA floodplains as well as the dates of digitization of the paper-based maps.

I study if digitization on its own facilitates access to risk information by focusing on communities which received a digital flood map depicting floodplain boundaries that are identical to the previously-available paper-based map. Whereas older flood maps were paper-based and hard to access, newer maps are digital and consolidated in an open online platform. I then investigate whether changing flood risk information impacts the demand for insurance by concentrating on communities where the new maps showed changes in the floodplain boundaries – either increasing or decreasing the number of properties located inside the high-risk zones. Finally, to quantify the accuracy of publicly-provided risk information and the welfare impacts of flood maps, I leverage discrepancies between the official flood maps and state-of-the-art estimates of flood risk within a model of insurance demand.

[1] The 100-year floodplain, also called the Special Flood Hazard Area (SFHA), is the area where the annual probability of flooding exceeds 1%. Flood maps are further described in Section 1.2 below.

Due to public funding constraints, the roll-out of digital flood maps was staggered in time and space, with different neighborhoods receiving updated flood maps in different years. This allows me to leverage difference-in-differences approaches to estimate the impacts of updates maps on the residential demand for insurance. Yet, credibly estimating the causal impacts of map updates on the demand for flood insurance is complicated by several factors. First, I show that the treatment date, i.e., the timing of reception of an updated (digital) flood map, was not random: communities that transitioned earlier from a paper-based to a digital flood map systematically differ from communities that transitioned later. Second, treatment *implementation* also differed in time: earlier map updates were more likely to digitize pre-existing floodplain boundaries, whereas later updates tended to rezone properties outside of the high-risk zones. To account for patterns of selection into treatment, I use heterogeneity-robust difference-in-difference estimators of the average treatment effect. In order to flexibly estimate the entire distribution of treatment effects and to account for variations in the number of properties rezoned inside and outside of the high-risk zones, I then leverage a novel approach based on the estimation of unit-specific and spatially clustered synthetic controls.

This analysis yields five main findings. First, I find that residential demand for insurance is sensitive to the information contained in the flood maps, but that digitization on its own does not matter: rezoning properties outside of the high-risk zones decreases insurance demand by more than 10% on average after two years while expansions of the high-risk zones cause an average increase in insurance take-up of 30% at the census tract-level. This highlights that households are attentive to publicly provided risk information, and confirms that demand for insurance increases with perceived risk. Yet, map updates rezoned more than one million properties *outside* of the high risk zones on aggregate, which lead to a decline in insurance coverage nationally. In contrast, map updates that purely digitized floodplain boundaries, without rezoning any properties relative to previously-available paper-based maps, have no detectable impact on demand for insurance. This suggests that the costs of accessing flood risk information do not substantially limit household's demand for insurance, but that publicly-provided flood risk information is a crucial determinant of household-level climate adaptation.

Second, I find that households form beliefs about their flood risk exposure beyond their own zoning in the flood maps: rezoning properties inside the high-risk zones causes a significant number of properties outside of the high-risk zones but located inside the same neighborhood to purchase insurance. Although some homeowners could theoretically respond to federal insurance requirements within the high-risk zones, these spatial spillovers indicate that beliefs about risk are informed by the flood maps and impact demand for insurance. Such spillovers are consistent with rational belief updating: while flood maps provide a *discrete* description of flood risk (either inside or outside of the high-risk zones), *actual* flood risk tends to vary continuously in space.

Third, I find that the impacts of flood maps were unequally distributed. Whereas households in wealthier areas appear more responsive to changing flood risk information, I find that map updates caused insurance demand to decline more in neighborhoods with a higher share of African Americans, a pattern primarily driven by the disproportionate rezoning of properties outside of the high-risk zones. While map updates tended to reduce insurance premiums in these neighborhoods, several of these communities were subsequently hit by hurricane Laura and hurricane Ida in 2020 and 2021. Flood map-induced declines in insurance coverage raise concerns regarding the vulnerability of these communities.

Fourth, comparing official flood map updates with independent estimates of flood risk reveals that map updates incorrectly reclassified more than one million properties, and that more than six million properties were incorrectly left outside or inside of the high risk zones during the mapping process – thus providing incorrect risk signals to more than seven million households. I discuss potential reasons for the disagreements between official risk estimates and the best available science. I estimate that even in neighborhoods where the updated flood maps appear incorrect for a majority of rezoned properties, households still respond to new floodplain boundaries. This suggests that publicly-provided flood maps may be the main source of information about flood risk.

Finally, to get back-of-the-envelope estimates of the welfare impacts of risk information I leverage the reclassifications of properties and the independent estimates of flood risk inside a structural model of insurance demand. I find that under plausible values of risk aversion parameters, the map updates *decreased* welfare by at least $300 million, spread across the income distribution. Correcting floodplain boundaries would yield gains exceeding 20 billion dollars annually, primarily in wealthy and majority-white neighborhoods.

Taken together, these results suggest that publicly-provided climate risk information is a key tool to promote private investments in residential flood insurance, and that a better characterization of flood risk through improved mapping could substantially increase insurance coverage nationwide. The costs associated with incorrect flood maps provide a cautionary tale regarding the provision of inaccurate information.

This paper most directly contribute to the literature focused on understanding low demand for flood insurance (Anderson, 1974; Michel-Kerjan, 2010; Petrolia et al., 2013; Atreya et al., 2015; Kousky et al., 2019; Mulder, 2019; Landry and Turner, 2020; Liao and Mulder, 2021). Several studies argued that observed low levels of demand for flood insurance cannot be rationalized by an intrinsically low willingness-to-pay, but instead may be due to information frictions and behavioral biases (Chivers and Flores, 2002; Kunreuther et al., 2013; Gallagher, 2014; Shao et al., 2017; Mulder, 2021; Wagner, 2021; Bradt et al., 2021; Hu, 2022). My results provide novel insights on the mechanisms that drive the demand for flood insurance and show that risk information, through the categorical definition of flood risk used in the NFIP (the location of properties inside versus outside of the high-risk zones) determines take-up. They further reveal why, despite costly investments in flood map updates, insurance take-up has remained low: map updates rezoned more than one million properties outside of the high-risk zones, which decreased aggregate demand for insurance.

This paper also contributes to the adaptation to climate change and natural disasters literature (Kousky, 2014; Hallegatte et al., 2018; Botzen et al., 2019; Kahn, 2021). Previous work found that flooding events did not systematically cause out-migration (Boustan et al., 2012; Hornbeck and Naidu, 2014; Chen et al., 2017; Graff Zivin et al., 2019), suggesting that mobility alone will not be sufficient to decrease communities vulnerability to natural disasters. On the other hand Turnham et al. (2011) and Nguyen and Noy (2020) provide evidence that flood insurance leads to increased community resilience, and a large body of work finds that correct insurance pricing is crucial to send signals that incentivize defensive investments (The World Bank and the Government of Mexico, 2012; Hudson et al., 2016; Barreca et al., 2016; Kousky et al., 2017; Davlasheridze et al., 2017). I find that by primarily rezoning properties outside of the high-risk zones and ignoring more than five million properties that should have been reclassified as "high risk", the previous flood map updates missed on the opportunity to provide accurate risk and price signals, thus jeopardizing household-level purchase of insurance and other defensive investments.

These findings further inform the ongoing policy debate regarding the funding of flood risk mapping. On September 14, 2021, the House Financial Services Committee approved provisions in a budget reconciliation bill that will provide an additional three billion dollars to fund flood map updates, more than the total funding allocated to flood risk mapping in the past seven years (House Financial Services Committee, 2021).[2] My results show that updating flood maps could substantially increase flood insurance coverage and welfare nation-wide, *but only if* these updates are carried out in agreement with the best available flood science.

This work further relates to the literature on the salience of public policies and the impact of government-provided information (Stigler, 1961; Kennedy et al., 1994; Konar and Cohen, 1997; Cutler et al., 2004; Chetty et al., 2009; Sallee, 2014; Cabral and Hoxby, 2015). I estimate that the mode of information provision (digital versus paper-based) does not matter in the context of flood risk. In contrast, I find that the information contained in the maps is salient to households, suggesting they are an effective policy tool to orient insurance purchase decisions. While previous work estimated the impact of flood maps on housing values, (Troy and Romm, 2004; Pope, 2008; Bin and Landry, 2013; Beltrán et al., 2018; Shr and Zipp, 2019; Gibson and Mullins, 2020; Hino and Burke, 2021), this paper is the first to systematically study the roll-out of flood maps and their impacts on the demand for insurance. The spillover effects I uncover further show that households pay attention not only to their own floodplain classification, but also to the classifications of other properties located within the same neighborhood. This implies that current hedonic estimates of the impacts of flood

[2] "In addition to amounts otherwise available, there is appropriated to the Administrator of the Federal Emergency Management Agency for fiscal year 2022, out of any money in the Treasury not otherwise appropriated, $3,000,000,000 to remain available until expended, necessary expenses for flood hazard mapping and risk analysis [...]"

maps on property values might be lower bounds on the real effects.

Finally, this work offers an innovative application of recent econometric methods used to estimate treatment effects under staggered treatment timing, and provides a novel approach to characterize treatment effect heterogeneity. Recent literature highlights major pitfalls of the standard panel fixed effects models when the policy is *staggered* and the treatment effects are heterogeneous between cohorts (Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfoeuille, 2020; Callaway and Sant'Anna, 2020; Sun and Abraham, 2021; Goodman-Bacon, 2021). The proposed solutions all rely on first estimating and then aggregating period and cohort-wise average treatment effects – an approach sensitive to the presence of outliers, and impossible to conduct to characterize heterogeneity when few units are treated at the same time. I adopt an approach based on the estimation of unit-specific synthetic controls to estimate the full distribution of treatment effects. While a growing literature focuses on the aggregation of synthetic controls to estimate unbiased average treatment effects (Acemoglu et al., 2016; Gobillon and Magnac, 2016; Robbins et al., 2017; Abadie and L'Hour, 2019; Ben-Michael et al., 2020; Abadie, 2020), the approach employed in this paper uses synthetic controls to investigate the distributional consequences of a staggered government policy while identifying the communities most impacted.

The remainder of this paper proceeds as follows. Section 1.2 provides additional background on the flood insurance rate maps. Section 1.3 presents the datasets used in this paper and highlights the main descriptive results. Section 1.4 presents estimates of the average effects of the flood map updates on the demand for insurance using heterogeneity-robust difference-in-differences estimators. Section 1.5 introduces local synthetic controls estimates and compares flood maps to a state-of-the-art model of flood risk. Section 1.6 presents a model of insurance demand to quantify the welfare impacts of flood map updates. Section 1.7 concludes.

## 1.2 Background on Flood Insurance Rate Maps

### 1.2.1 Policy context

Flood risk is hard estimate – like other "rare" events, reliance on historical data to predict occurrence probabilities suffer from a small sample issue. The inability to predict flood events and damages lead to the unraveling of the private flood insurance market following the 1927 Mississippi River flood. Since its inception in 1968, the functioning of the National Flood Insurance Program crucially relies on the provision and updating of Flood Insurance Rate Maps (FIRMs, colloquially called "flood maps") with the dual goals of communicating flood risk assessments and setting federally-provided flood insurance premiums. These maps are provided at the community level, where communities can be towns, cities or counties (see Knowles and Kunreuther (2014) for a discussion of the National Flood Insurance Program's history).

While flood risk is hard to assess, communicating flood risk to non-experts is an additional challenge. Flood risk involves a property-specific probability distribution function over inundation depths, where costly events are typically found in the upper-tail of the distribution. In order to convey flood risk probabilities to the general public and urban planners alike, the NFIP's flood maps simplify underlying risk calculations to provide estimates of flood risks that are differentiated by *zones*. In particular, flood maps highlight the 100-year floodplain, also called Special Flood Hazard Area (SFHA) or high-risk zone: in this area, the annual probability of a flood event exceeds 1% per year.[3] Flood maps also highlight the 500-year floodplain (or X shaded zones), where the annual probability of flooding is estimated to be between 0.2% and 1%. Areas outside of these zones are classified as exposed to "minimal flood risk".

In principle, flood zones could be defined for any annual probability of flooding.[4] In practice, since their initial definition, the 100-year floodplain plays an essential role: (i) flood insurance is mandatory for properties with a federally-backed mortgage that are located within the 100-year floodplain, (ii) flood insurance premiums vary discontinuously at the 100-year floodplain borders, and (iii) urban planning often restricts development within the 100-year floodplains.[5]

Until the 2000s, almost all flood maps were paper-based. These maps were hard to find, hard to read for non-experts, and generally outdated. An early effort was made between 1995 and 2005 to partially digitize flood maps (also known as the Q3 data product, presented below), but it was only viewed as a temporary substitute for the required in-depth modernization of the paper-based maps (Gall et al., 2007), and as such was not widely distributed to the public.

In order to improve flood risk identification and to facilitate information access, a presidential initiative called the Map Modernization program (MapMod) allocated a budget of approximately 1.4 billion USD starting in 2005 to update paper-based FIRMs to digital flood insurance rate maps (DFIRMs, or simply "digital maps"). This made MapMod one of the most expensive publicly funded mapping programs in history (see Morrissey (2008) for detailed background on the program). In 2005, due to scarce funding, Congress urged FEMA to target the modernization of flood maps in areas most vulnerable to flood risk and densely populated, with the goal to cover 92% of the population with digital maps by the end of the program). Although some FIRMs were transitioned to DFIRMs in 2005, more than 45,000 census tracts still

[3]The SFHA is further divided into V zones, which depicts areas subject to breaking waves of at least 3 feet, and A zones.
[4]For instance, the flood maps could highlight the 200-year and the 1000-year floodplains within the current X zones
[5]This has spurred criticism of the 100-year floodplain concept as both an insurance-setting and a communication tool from academics and practitioners alike (Wittenberg, 2017; Koerth, 2017; Bell and Tobin, 2007; Ludy and Kondolf, 2012; Highfield et al., 2013).

did not have a digital map by the end of the MapMod program in 2008.[6] In 2008 the MapMod program was rolled into the Risk Mapping, Assessment, and Planning program, which essentially continued to modernize flood maps with revised objectives to cover 100% of the population with valid digital flood data (FEMA, 2009). The total costs of the Risk Mapping, Assessment, and Planning program were estimated to four billion dollars (Department of Homeland Security, 2012).

To this day, funding is continuously allocated to update the remaining paper-based maps to digital maps: in late 2019, 16 million residential properties, or about 12% of the residential housing stock in the contiguous US, were still not covered by a digital map.[7] In addition, because flood risk and flood risk assessment technologies both evolve, the Federal Emergency Management Agencies (FEMA) is requiered to update flood maps at least every five years – whether or not these maps are already digital. Yet, due to funding and capacity constraints, few communities ever received a second digital map. Among the communities that did, the map update itself took an average of three to five years (Horn and Webel, 2019; Wilson and Kousky, 2019).

The flood maps digital transition reduced the cost of accessing flood risk information and, in many cases, changed the underlying floodplain boundaries. To appreciate how digitization improved access to information, it is useful to walk through a specific example. Imagine a resident living in Upper Merion Township, Pennsylvania, wanted to learn about her property's exposure to flood risk before the Montgomery County flood map had been digitized.[8] Because flood maps are divided into "panels", or rectangular sections of the U.S., she would first have to figure out which panel number includes her property: this implies finding the Montgomery Map Index and identifying the relevant panel code (see Figure 1.1.A). She would then use this code to find the relevant paper-based map and attempt to locate her property, where the 100-year floodplain is represented in dark grey and the 500-year floodplain in light grey. In contrast, since the updated digital flood map (effective in 2016), she can access the National Flood Hazard Layer, a publicly available online platform, and directly locate her property by typing her address in the search bar(see Figure 1.1.C). The 100-year floodplain is now depicted in blue, and the 500-year floodplain is represented in orange.[9]

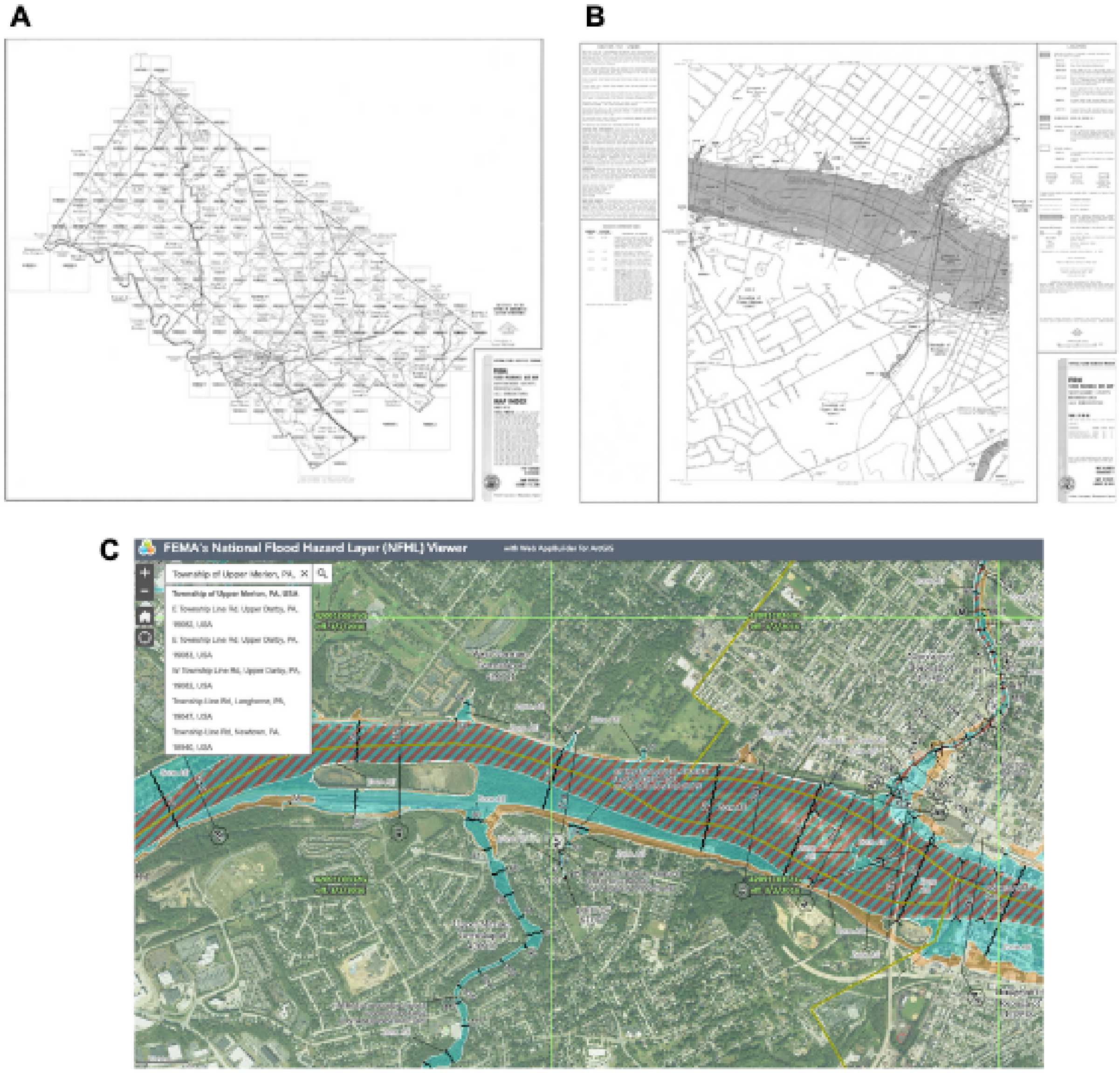
[6]The initial goal of the Map Mod program (2003-2008) was to convert all paper-based FIRMs to updated, digital flood maps. However, policy analysts and practitioners alike realized as early as 2005 that there was a need to target scarce funding resources. This was formalized as a revision to the Map Mod objectives called the "Mid-Course Adjustment" (FEMA, 2006). It is also during the Mid-Course Adjustment that the modernization started using census-block data instead of county-wide panels (Morrissey, 2008)

[7]Author's calculations, shown in Table 1.3.

[8]Hurricane Ida severely hit Montgomery County in September 2021, causing the deaths of at least three people.

[9]Red and blue stripes represent the regulatory floodway).

Figure 1.1: Evolution of the Flood Insurance Rate Maps in Montgomery County, PA




**A**: Map Index from Montgomery County, PA, effective in 1999. **B**: FIRM number 42091C0351F, effective in 1999. The 100-year and 500-year floodplains are shown in dark and light grey, respectively. **C**: Screenshot of the DFIRM effective in 2016, as shown on the National Flood Hazard Layer. The 100-year floodplain is depicted in blue and blue/red stripes, while the 500-year floodplain is depicted in orange. Sources: FEMA's Map Service Center and National Flood Hazard Layer.

Figure 1.1 shows a situation where floodplain boundaries did not substantially change through the map update. This is not always the case: map updates can rezone properties either outside or inside of the 100-year and 500-year floodplains. Expansions are likely to occur due to increased flood risk caused by a changing climate, but reductions can be warranted if the earlier maps were based on outdated flood science or did not account for newly constructed protective structures such as levees.[10] In the rest of the paper, I use pure digitizations as well as changes in floodplain boundaries to assess the impacts of both access to information and of changing risk information on the demand for flood insurance.

### 1.2.2 Stylized model of the impact of flood map updates on the demand for insurance

A priori, the impact of map updates on the demand for flood insurance is ambiguous. To fix ideas, it is useful to walk through a simple model. Let the (uncompensated) demand for flood insurance be defined as

$$x \equiv x(b_f(\nu), p(\nu), m(\nu)) \tag{1.1}$$

where $\nu$ is the risk shown on the flood map, $b_f$ is the individual's *belief* about the probability of flooding, $p$ is the price of insurance relative to a composite good, m denotes the role of the mandate on insurance demand.

The impact on demand of a flood map update that depicts a marginal increase in flood risk can be decomposed as follows:

$$\frac{dx}{d\nu} = \underbrace{\overbrace{\frac{\partial x}{\partial b_f}}^{\geq 0}\overbrace{\frac{\partial b_f}{\partial \nu}}^{\geq 0}}_{\geq 0} + \underbrace{\overbrace{\frac{\partial x}{\partial p}}^{<0}\overbrace{\frac{\partial p}{\partial \nu}}^{>0}}_{<0} + \underbrace{\overbrace{\frac{\partial x}{\partial m}}^{\geq 0}\overbrace{\frac{\partial m}{\partial \nu}}^{\geq 0}}_{\geq 0} \tag{1.2}$$

If the individual is not attentive to flood maps, then $\frac{\partial b_f}{\partial \nu} = 0$, otherwise this term is positive. In general for a risk averse individual, the increased beliefs about flood risk will (weakly) increase demand for insurance, and $\frac{\partial x}{\partial b_f}$ will be positive.[11] In addition, an updated flood map showing increased flood risk will in general increase insurance premiums, and the second term will be negative.[12] Finally, assuming the increase in flood risk is enough to impact insurance mandate requirements, the last term will be non negative. However, in the context of flood insurance, previous surveys of property owners and recent work argue that most households are unaware of the mandate requirement inside the 100-year floodplain, i.e. $\frac{\partial m}{\partial \nu} = 0$, and that even if people were aware, the mandate is not enforced, i.e. $\frac{\partial x}{\partial m} = 0$ (Tobin and Calfee, 2005; Michel-Kerjan et al., 2012; National Reearch Council, 2015; Wagner, 2021). A lack of either awareness or enforcement would imply that the last term in equation 1.2 is equal to 0. Yet, even in this restricted setting, the impact of an increase in (nominal) risk on the demand for insurance is ambiguous.

[10] This last point is especially contentious, as levees are not always properly maintained (Ludy and Kondolf, 2012).

[11] In the limiting case where pricing is fair and the individual is a risk averse expected utility maximizer, the effect of an increase in risk on demand will be strictly positive at $b_f = 0$, and equal to zero at any strictly positive value of $b_f$.

[12] I assume that insurance is a normal good.

The digitization of paper-based maps, without any changes in floodplain boundaries, only impacts the first term. In this scenario, a property owner who lives inside the floodplain and initially believed she was not at risk of flooding could learn something new about her risk exposure due to digitization alone – as long as the change in flood risk depicted on the flood maps is salient to her ($\frac{\partial b_f}{\partial \nu} \geq 0$). A risk averse consumer would then purchase insurance. The next section describes the approach used to separately estimate the impacts of flood maps digitization and flood map rezoning on the demand for flood insurance.

## 1.3 Data and descriptive statistics

### 1.3.1 Digital flood insurance rate maps

I collect and compile data on all digital flood maps released between 2005 and 2019, obtained both through a Freedom of Information Act request and following meetings with former GIS analysts who performed contractual work for FEMA. These digital maps provide both the *date* when the DFIRM became active, and the *spatial polygons* of the 100-year and 500-year floodplains.[13] Figure 1.2.A depicts the cumulative number of census tracts that received a DFIRM since 2000. Almost all census tracts are now covered by a digital map, with the greatest number of yearly digitizations occurring since 2007. Although flood maps are supposed to be updated at least once every five years, to date less than 20% of census tracts have received more than one DFIRM, covering only 16 million residential properties, or less than 10% of the residential housing stock (see Table B.2 in the appendix). The main econometric analysis of this paper focuses on census tracts that receive their first digital flood map, as this allows me to estimate the impact of digitization.[14]

Figure 1.2.B shows the spatial distribution of the census tracts that received at least one digital flood map and the year during which they received it. More than 85% of census tracts are now covered by a digital map, with most states showing within-state variation in the timing of digitization, although there are some exceptions (North Carolina, for instance, had a separate floodplain mapping program which started to leverage digital technologies to produce the flood maps as early as 2004).

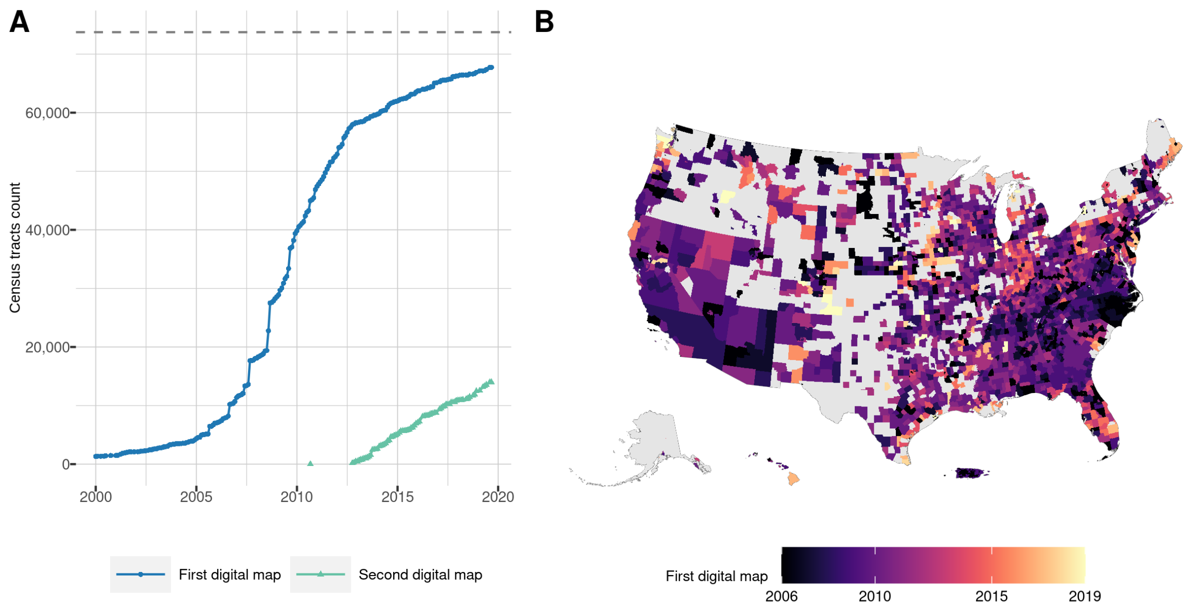
### 1.3.2 Digitized version of paper-based flood maps: the Q3 data

Most flood maps were paper-based prior to 2007. Because we are interested in how the modernization of flood maps process changed the spatial extents of the floodplains, we need a GIS-ready version of these paper-

[13] The steps used to process these maps are provided in Appendix A.0.1.

[14] This further facilitates the definition and interpretation of treatment effects: for census tracts that received more than one digital flood map, defining treatment status after the first digital map and before the second map is ambiguous. Resolving the ambiguity requires to either (i) make assumptions about the dynamics of the treatment effects and leverage an ad hoc threshold above which the first treatment effect is considered to have vanished, or (ii) define treatment as a qualitative variable taking more than two values, which further complicates the estimation of treatment effects.

Figure 1.2: Flood maps digitization throughout the U.S.




**A**: Cumulative number of census tracts that received a first digital flood map (in blue circles) or a second digital flood map (green triangles). The grey dotted line shows the total number of census tracts in the U.S. (73,745 tracts, including Puerto Rico). **B**: Spatial release of the first digital flood maps, censored below in year 2006 for clarity.

based maps. Luckily, FEMA produced digitized versions of the 100-year and 500-year floodplain boundaries depicted on the paper-based maps that were effective in over 1,300 counties at the time of scanning (1996-2000). This product, known as the Q3 data, makes it possible to assess how FEMA's floodplains evolved over time (FEMA, 1996).[15] Although I can only track changes in the floodplains for which Q3 data exist, this encompasses more than 48,000 census tracts and includes the most flood-prone areas, where the majority of insurance policies are purchased.[16]

Figure 1.3.A shows the digital version of the 1984 flood map in New Orleans,[17] while Figure 1.3.B shows the digital flood map that became effective in September 2016. In both digital products, the 100-year floodplain is depicted in light blue while the 500-year floodplain is shown in orange. Figure 1.3.B also depicts pink areas for the zones "protected by levees" (outside of the floodplains). In New Orleans, the map modernization program led to a massive rezoning of properties *outside* of the SFHA, and inside both the 500-year floodplain or outside the floodplains altogether. This rezoning was rationalized by substantial

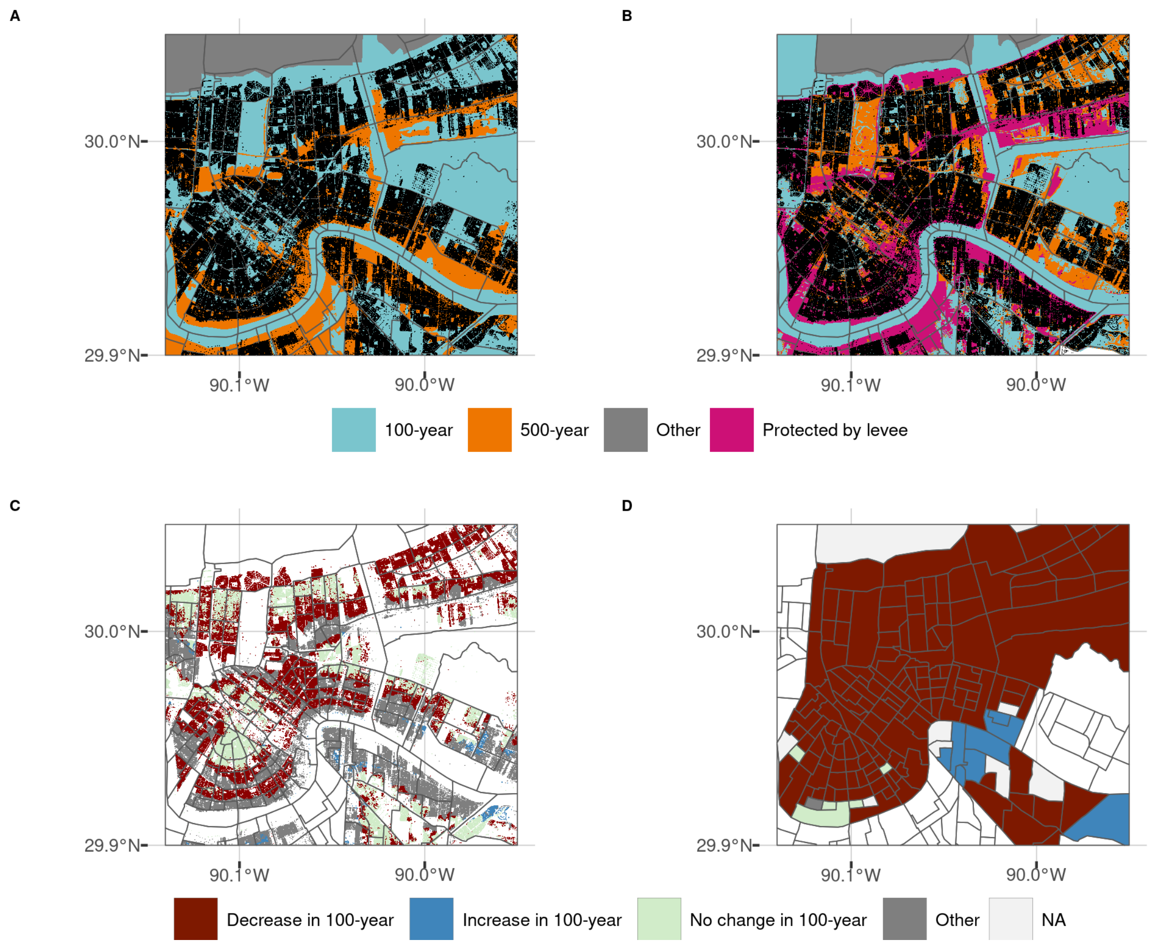
[15] It was already understood in the 90s that providing fully digital FIRMs would require several years as well as substantial financial investments. The Q3 data was meant as an early digitization effort to quickly distribute the floodplain data to practitioners in order to facilitate disaster recovery and planning activities. It became the standard product for planning under the Disaster Mitigation Act of 2000 (Gall et al., 2007)

[16] Out of the 2,814,354 residential flood insurance policies active in January 2009, 85 % were in areas covered by a Q3 map.

[17] The original paper-based map, digitized without changes in the underlying boundaries in the Q3 product, is presented in the Appendix Figure A.6.

investments made by the U.S. Army Corps of Engineers in flood control systems such as levees, floodwalls, floodgates and pump stations (FEMA, 2016).

Figure 1.3: Changes in the Flood Insurance Rate Maps in New Orleans, Orleans Parish, LA




**A**: Q3 data product showing the flood zones effective in 1984, combined with geolocalized residential properties in Orleans Parish (black dots). **B**: Current DFIRM showing the flood zones effective as of September 2016, combined with geolocalized residential properties. **C**: Changes in the 100-year floodplain computed at the property level. **D**: Changes in the number of residential properties in the 100-year floodplain between 1984 and 2016, aggregated at the census tract-level and classified into three categories (increase, decrease, or no change).

### 1.3.3 Population located within the different floodplains

To quantify how many properties are affected by the map modernization program within each census tract, I intersect the Q3 data and digital flood maps with a dataset of geolocalized residential properties provided by the First Street Foundation. This dataset consolidates records from county assessor offices and provides the coordinates of more than 125 million residential properties in the contiguous United States. In the census tracts that are covered by Q3 data and which have received a digital flood map, I compute whether each property is zoned inside or outside of either the 100-year and 500-year floodplains within each data product

(Figure 1.3.C). For the empirical estimation of section 1.4 below, I then aggregate these changes at the census tract-level and classify census tracts depending on whether the map update rezoned more than 1% of residential properties properties inside or outside the 100-year floodplain on aggregate, or whether the map update did not change the 100-year classification of any property (Figure 1.3.D).[18]

Figure 1.4 depicts how the modernization of flood maps impacted the number of properties classified inside the 100-year floodplain at the census tract-level, across the contiguous US.[19] The states in which significant digitization of the FIRMs occurred all have communities where the map updates both increased and decreased the number of properties located in the high-risk zones. However, when aggregating these changes, the total number of properties located within the 100-year floodplain *declined* by 1.1 million. A few states are responsible for the exclusion of most properties: on aggregate, Florida and California rezoned 450,000 and 250,000 residential properties residential outside of the 100-year floodplain, respectively.[20]

This aggregate removal of properties from the high-risk flood zones is puzzling. There is an overwhelming consensus among scientists that flood risk in the US is increasing, which everything else equal should lead to expansions of the 100-year floodplain (Marsooli et al., 2019; Bates et al., 2021; Wing et al., 2022). Three different reasons could explain the observed reductions. First, older paper-based flood maps could be systematically incorrect. Flood risk is incredibly hard to assess, and the first flood maps relied on coarse models of flood risk which could have systematically overestimated the extent of the 100-year floodplain. Second, over the past two decades substantial investments were made in flood defense systems, such as in New Orleans. Accounting for these adaptation infrastructures could warrant reductions in the 100-year floodplain. Finally, some evidence and field studies suggest that local homeowners actively lobby against expansions of the 100-year floodplain, as these expansions would increase flood insurance premiums and decrease property values (Pralle, 2019; Lea and Pralle, 2022). Although I cannot disentangle these mechanisms, in section 1.5.5 below I provide evidence suggesting that only a portion of the reductions of the 100-year floodplain were warranted, while many more properties were incorrectly left out of the high-risk zones during the map updates.[21]

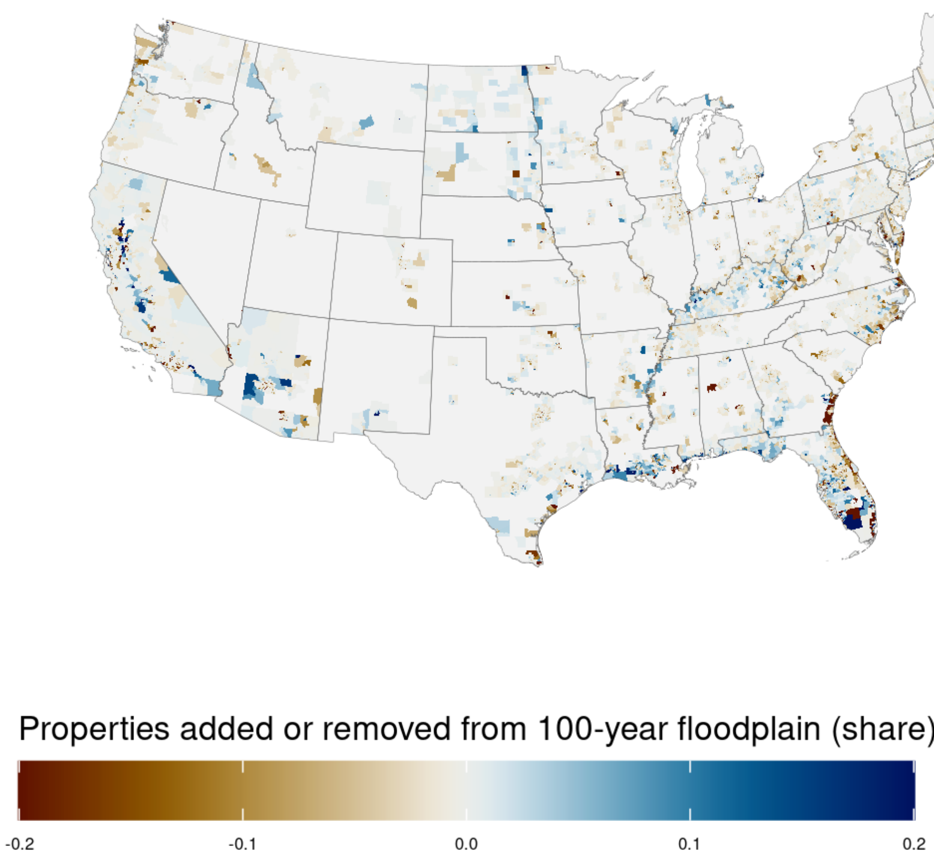
[18] In the Appendix Figure A.4, I show a similar procedure for the South-Eastern part of Broward County (FL), an area also subject to substantial flood risk.

[19] Unfortunately, the data are currently missing a large number of geolocalized properties in Alaska, Hawaii and Puerto Rico, which are omitted from the remaining analysis.

[20] To assess the representativeness of the First Street Foundation data to construct measures of floodplain exposure, I use two alternative measures of population zoned in different floodplains based on i) EPA's EnviroAtlas dasymetric reallocation layers and ii) the ZTRAX data. Both measures provide conclusions in agreement with the FSF properties-based measure. See section A.0.3 in the Appendix for additional details.

[21] Table B.2 in the appendix shows that among the places that received two digital flood maps, the second map update also removed properties from the 100-yeaar floodplain on aggregate.

Figure 1.4: Change in residential properties located in the 100-year floodplain during the map update




Changes in the number of properties zoned inside the 100-year floodplain are expressed relative to the total number of residential properties in each census tract. The figure depicts changes bounded between -0.2 and 0.2 to improve color contrasts.

### 1.3.4 Flood insurance policies

I use data on flood insurance policies provided by FEMA covering all NFIP policies active between 2008 and 2019 obtained through the OpenFEMA platform and following three distinct Freedom of Information Act requests. Each of the 60 millions observations is an individual insurance policy with information on when the policy was active, details about the insured property such as the number of floors and the FEMA flood zone in which the property is located, as well as details about the insurance policy itself (coverage and deductible amounts for both building and contents, and total policy cost). Importantly, the dataset does *not* include property identifiers, and policies are not geocoded. This means that I cannot track changes in insurance demand at the property-level. For the analyses below, I construct a panel of active policies at the census tract / month level. Details about the construction of the panel are provided in Appendix A, and a chart summarizing the different datasets used in this paper is provided in Figure A.5.

The upper part of Table 1.1 presents some summary statistics on the panel of insurance policies. There is a wide range in the number of active policies per census tract, varying from 0 to more than 6,000. This spread will partially motivate the estimation strategy I conduct at the census tract-level in section 1.5 below. In some tracts, all of the policies are covering properties located inside the high-risk zones, whereas in others only properties located outside of the 100-year floodplain are covered by flood insurance policies. The zoning inside and outside of the 100-year floodplain explains a substantial share of the variation in average insurance premiums observed between tracts: the average policy cost is about \$596 per year, but is \$419 outside of the 100-year floodplain while it is \$1,012 inside of it.

Table 1.1: Summary statistics, estimation panel and census tracts cross section

| Variable | N | mean | sd | min | max |
|---|---|---|---|---|---|
| **Panel**: | | | | | |
| Number of active policies | 7056060 | 59.21 | 190.97 | 0 | 6415 |
| Number of active policies inside 100-y | 7056060 | 31.29 | 140.82 | 0 | 5322 |
| Number of active policies outside 100-y | 7056060 | 27.92 | 101.03 | 0 | 5314 |
| Share policies inside 100-y | 6333679 | 0.33 | 0.33 | 0 | 1 |
| Average policy cost | 6369408 | 596.04 | 347.86 | 6 | 10143 |
| Average policy cost inside 100-y | 4095644 | 1012.87 | 629.77 | 42 | 11665 |
| Average policy cost outside 100-y | 6110812 | 418.74 | 155.64 | 6 | 3422 |
| Average construction of insured property | 6369408 | 1972.23 | 20.31 | 1900 | 2018 |
| Average initial insurance year | 6369408 | 2008.49 | 3.77 | 1980 | 2020 |
| Average coverage | 7056060 | 192795.4 | 103611.54 | 0 | 7173600 |
| Cost per thousand dollar insured | 6369408 | 3.05 | 2.52 | 0.54 | 173.04 |
| Cost per thousand dollar insured inside 100-y | 4095644 | 6.6 | 4.07 | 0.15 | 420 |
| Cost per thousand dollar insured outside 100-y | 6110812 | 1.77 | 1.19 | 0.48 | 36.1 |
| **Cross section**: | | | | | |
| Year of map modernization | 65341 | 2009.24 | 3.72 | 1995 | 2019 |
| Disaster declaration within two years of treatment | 65314 | 0.54 | 0.5 | 0 | 1 |
| Has a Q3 map | 71183 | 0.76 | 0.43 | 0 | 1 |
| Has a digital map in 2019 | 71183 | 0.92 | 0.27 | 0 | 1 |
| Relative change in properties zoned inside 100-y | 49476 | -0.02 | 0.12 | -1 | 1 |
| Share population Black or African American | 70801 | 0.14 | 0.22 | 0 | 1 |
| Median household income, past twelve months | 70440 | 64314.44 | 32176.82 | 2499 | 250001 |

## 1.4 Estimating the average effects of flood map updates

### 1.4.1 Selection into treatment timing and implementation

While FEMA started to actively modernize flood maps in 2006, the *timing* and *implementation* of the map updates were not random.

The Map Modernization and then Risk Mapping programs had concurrent objectives of providing digital flood maps covering the largest population possible and in areas most vulnerable to flood risk. These objectives can conceivably conflict with one another. Table 1.2 offers a regression-based summary of how these conflicts were resolved. In the first two columns, the dependent variable is the year in which the census tract receives its first digital flood map, while in columns 3 and 4 the dependent variable is the net share of residential properties in the census tract that were rezoned inside the 100-year floodplain on the updated map (counting rezoning outside of the 100-year floodplain as negative). All models include fixed effects for each FEMA Region, as the roll-out of digital flood maps is primarily decentralized at this level.

Table 1.2: Selection into treatment timing and implementation

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Treatment year | | Share rezoned inside 100-year f.p. | |
| | (1) | (2) | (3) | (4) |
| Population density (IHS) | $-0.201^{***}$ (0.066) | $-0.277^{***}$ (0.056) | $-0.009^{**}$ (0.004) | $-0.003^{***}$ (0.001) |
| Median income (IHS) | $-0.727^{**}$ (0.289) | $-0.807^{***}$ (0.285) | $-0.009$ (0.007) | $-0.001$ (0.003) |
| Share African Americans (IHS) | 0.120 (0.721) | 0.276 (0.697) | $-0.007$ (0.020) | $-0.032^{*}$ (0.017) |
| Disaster declaration prior to treatment | | | 0.027 (0.022) | 0.016 (0.013) |
| Average Annual Loss (IHS) | 0.095 (0.059) | | $-0.002^{*}$ (0.001) | |
| Insurance policies/property, 2008 (IHS) | | $4.733^{***}$ (1.455) | | $-0.488^{***}$ (0.149) |
| Treatment year | | | $-0.004^{**}$ (0.002) | $-0.002^{**}$ (0.001) |
| Fixed Effects | FEMA Region | FEMA Region | FEMA Region | FEMA Region |
| Mean outcome | 2009.2 | 2009.2 | -0.018 | -0.017 |
| Observations | 48,944 | 49,181 | 34,631 | 34,631 |
| $R^2$ | 0.087 | 0.103 | 0.057 | 0.218 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
All independent variables are transformed using the inverse hyperbolic sine
Standard errors clustered at the county level

Columns 1 and 2 reveal that census tracts with greater population density received a digital flood map earlier on average: a 1% increase in census tract density is associated with a treatment date that is between 0.1 and 0.2 years earlier, consistent with some targeting of populous communities in order to comply with policy mandates and deadlines.

To study whether areas more vulnerable to flooding were more likely to receive an updated flood map earlier, I use a measure of predicted average annual economic loss per property produced by the First Street Foundation (column 1). Interestingly, the point estimate is *positive*, suggesting that more flood-prone areas were treated *later*, although the effect is small and noisy. I also estimate a model that captures flood vulnerability through the share of residential properties within a census tract that are covered by flood insurance in 2008 (column 2). This proxy for flood vulnerability is not ideal, as the next sections are devoted to showing that digital flood maps have large impacts on insurance take-up. But given that insurance take-up is a metric directly observable by policymakers, it potentially provides valuable information about where flood maps might be most used. Contrary to official policy guidelines, places that had higher flood insurance coverage in 2008 were more likely to receive their digital flood map *later*: a ten percent increase in the rate of residential property insurance coverage is associated with a treatment timing delayed by more than .4 years. Although the previous estimates are correlational, they suggest that cohorts treated in different years might respond differently to treatment.

To assess the extent of heterogeneity in policy implementation (the *intensity* of treatment), I now focus on the rezoning of properties inside or outside of the 100-year floodplain (columns 3 and 4 in table 1.2). As the rest of this paper will show, the zoning of properties inside the 100-year floodplain has substantial impacts on insurance take-up. Strikingly, census tracts with higher predicted flood losses (column 3) or greater insurance coverage in 2008 (column 4) experienced *less* rezoning inside the 100-year floodplain. Census tracts that were treated earlier also saw less rezoning of properties inside the 100-year flooplain: after accounting for the other census tracts characteristics, a one-year increase in treatment year implies on average a -0.004 decline in the share of properties rezoned inside the 100-year floodplain (about 20% of the outcome's average value in the sample). Table B.3 in the appendix shows these main conclusions hold when restricting the estimation to the sample of census tracts that are observed with at least one active insurance policies at all times.

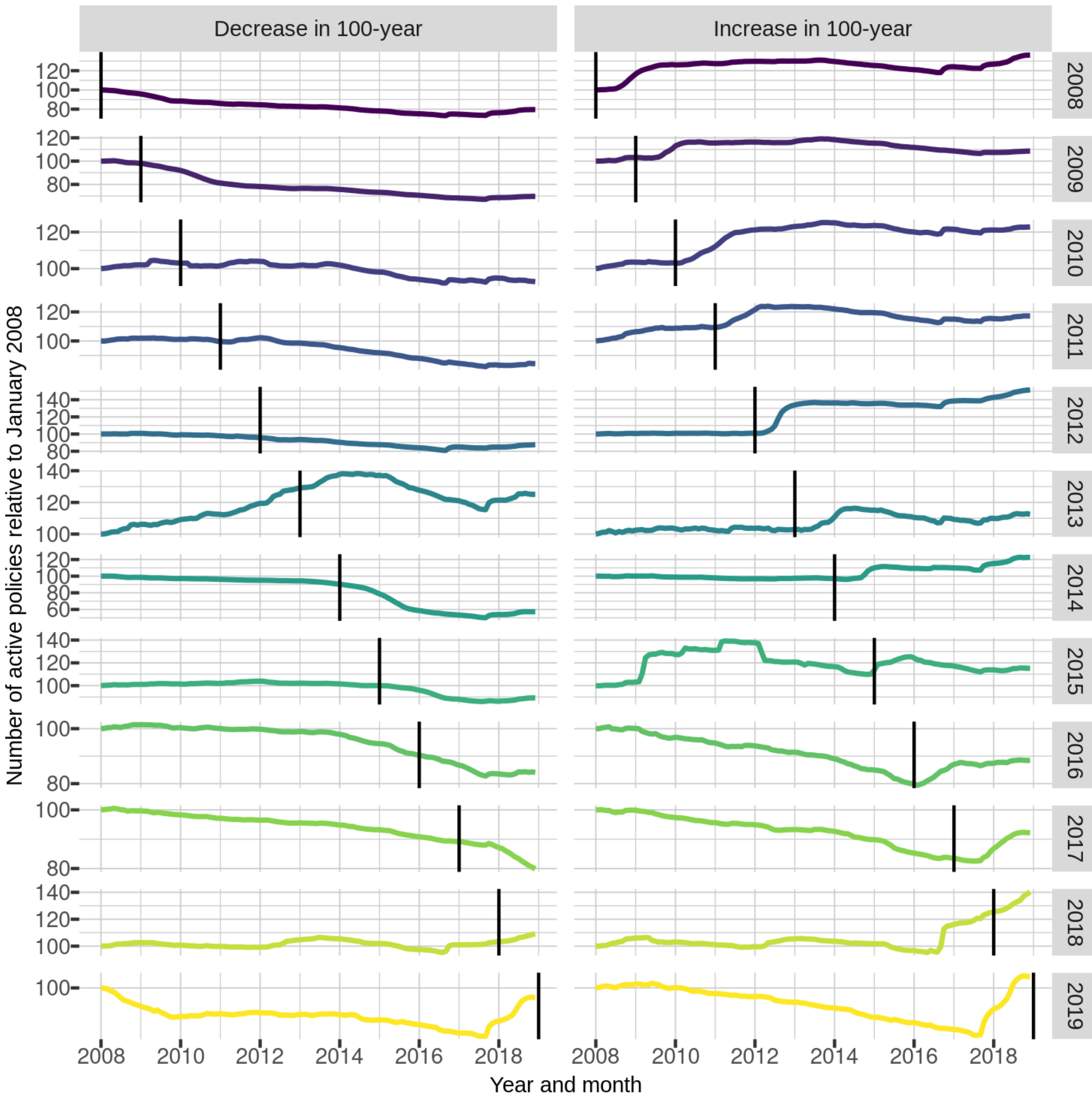
Overall, the previous results show that flood maps were first modernized in areas with higher population density, consistent with the policy mandate to cover most of the nation's population with digital flood mapping products, but also in areas less vulnerable to flood risk, which is in contradiction with the policy mandate to focus on flood-prone areas. Although I cannot provide credibly causal evidence on what caused the delay in flood map updates in areas more flood-prone, anecdotal evidence and discussions with floodplain managers suggest that it was due to (i) the complexity of modelling flood risk in these areas, and (ii) local homeowners lobbying against new flood maps. Accounting for these patterns of selection into treatment directly informs the estimation strategies I employ below.

### 1.4.2 Raw evidence on the impacts of flood map updates

In order to further motivate the econometric models, it is instructive to look at the patterns of insurance demand already visible in the raw data. Figure 1.5 depicts the number of insurance policies covering residential properties in each month relative to January 2008. Insurance policies are depicted separately by year of the flood map update (vertical facets) and based on whether the map update rezoned properties outside (left) or inside (right) the 100-year floodplain on net at the census tract level.

Whatever the year in which the flood map update occurs, following the updates the demand for flood insurance appears to decrease where the map updates rezoned properties outside of the 100-year floodplain (left facets, with the exception of 2013), while the demand for insurance increases in census tracts where

Figure 1.5: Raw evidence on the impact of floodplain rezoning on the demand for insurance




Number of active flood insurance policies covering residential properties in each month relative to January, 2008 (baseline 100). The number of insurance policies are depicted separately by year of the flood map updates (vertical facets) and based on whether the map updates rezoned properties outside (left) or inside (right) the 100-year floodplain at the census tract level. Black vertical lines show the year of the map update.

the map updates rezoned properties inside the 100-year floodplain (right facets). While only suggestive, these patterns indicate that map updates may have substantial impacts on the demand for flood insurance, depending on the rezoning of properties inside and outside of the 100-year floodplain.

### 1.4.3 Heterogeneity-robust event studies

Our empirical setting involves a large number of units (census tracts) that get treated (receive a digital map) in different time periods (year and month). In such contexts, where the treatment is *staggered*, recent econometric work highlights important pitfalls with the use of the two-way fixed effect estimators that is still standard in the applied literature. In particular, de Chaisemartin and D'Haultfoeuille (2020) demonstrate that if the treatment effect is heterogeneous between *cohorts* (group of units treated simultaneously), then the two-way fixed effect estimator does not recover the Average Treatment Effect on the Treated (ATT).[22] Sun and Abraham (2021) show that even the more flexible event-study specification used to estimate dynamic effects is not robust to heterogeneity between cohorts, and they provide a general formula for the bias that arises in various event-studies specification.[23] Recent applied work shows such pitfalls make two-way fixed effect estimators extremely sensitive to minor specifications of the regression model (Weill et al., 2021).

In our context, treatment effects are very much likely to be heterogeneous between cohorts. Section 1.4.1 above shows that census tracts treated earlier are systematically different from census tracts treated later (potential heterogeneity due to *selection*), and census tracts treated later received an updated map that tended to rezone more properties outside of the 100-year floodplain (potential heterogeneity due to varying *implementation*). In addition, the impact of a new map on neighborhood-level insurance demand will *mechanically* be dynamic, because flood insurance policies are usually purchased for an entire year. Few individuals cancel their policy before its expiration date, since cancellations do not automatically lead to reimbursements.[24] Finally, purchases of insurance policies are not uniform within a year: more policies are purchased in the summer months. These aspects lead me to consider heterogeneity-robust estimators of the dynamic treatment effects.

I follow the approach proposed by Callaway and Sant'Anna (2020) and estimate a distinct treatment effect for each cohort and period relative to treatment. To understand how within-tract rezoning impacts

[22]This is true even when the identifying "parallel trends" assumption holds. **?** provides an intuitive decomposition of the two-way fixed effect estimator as a weighted average of all possible two-by-two diff-in-diff estimators, where units treated "early" serve as implicit controls for units treated "later", which can be be especially problematic if the treatment effect is dynamic.

[23]They further highlight that the estimated effects on leads and lags of the treatment can be contaminated by other time periods *even if* treatment effect is homogeneous.

[24]For instance, consider a homeowner who purchased flood insurance exactly one month before the new digital flood map is released, and the new flood map induces her to stop buying flood insurance. Because cancelling a flood insurance policy is costly, this homeowner may "passively decide" to remain insured for up to 11 months after the new flood map was released, and then wait until the end of the one year term to decide to not renew her policy. In this scenario, the individual decision of interest is the one to not renew the policy.

the demand for insurance, I further separate treatment effect estimates by type of rezoning that occurred based on rezoning inside or outside the 100-year floodplain (on net) at the tract level.[25] Formally, we are interested in the Average Treatment effects on the Treated (ATTs) for each group $g$, time period $t$, and net flood map rezoning at the tract level $x$:

$$ATT(g,t,x) = \mathbb{E}(Y_t(g,x) - Y_t(0,x)|G_g = 1, X = x) \tag{1.3}$$

where $Y_t(g,x)$ denotes the potential outcome of units in group $x$ at time $t$, *if* they were to become treated at time $g$. $Y_t(0,x)$ denotes the potential outcome of these same units had they not received treatment. For each unit we can only observe the *realized* outcome: to estimate the ATT, the expectation over potential outcomes is replaced with sample averages, where the control groups are made of units that have not yet received treatment at time $t$. I construct the groups $x$ to reflect treatment intensity based on how the modernization of the flood maps impacted rezoning into the 100-year floodplain. Specifically, $x$ is one of {Increase in 100-year floodplain, Decrease in 100-year floodplain, No change in 100-year floodplain}. As before, increase or decrease in the 100-year floodplain denote tracts where more (less) than 1% of residential properties are rezoned inside (outside) of the 100-year floodplain on net respectively, while "No change in 100-year floodplain" denotes tracts where there was no rezoning of any property relative to the 100-year floodplain. Note that we could consider other groups, for instance the groups made of census tracts where the new maps rezoned more than 2, 3 or 5 % of properties inside or outside of the floodplain. The issue with such approach to heterogeneity is that some of these groups will contain very few units, making hard (and in many cases impossible) to estimate average treatment effects. This issue is tackled in section 1.5.1 below.

Once cohort-specific ATTs are estimated, I aggregate them into ATTs by length of exposure using cohort sample-size weights:

$$\theta(e,x) = \sum_g \mathbf{1}\{g + e \leq T\} \cdot P(G = g \ \cap \ X = x|G + e \leq T\} \cdot ATT(g, g + e, x) \tag{1.4}$$

where $e$ is the exposure time ($e = t - g$), g is an index for the cohort of flood map update, G is the time period that a unit is first treated, T is the last time period for which ATTs are identified.

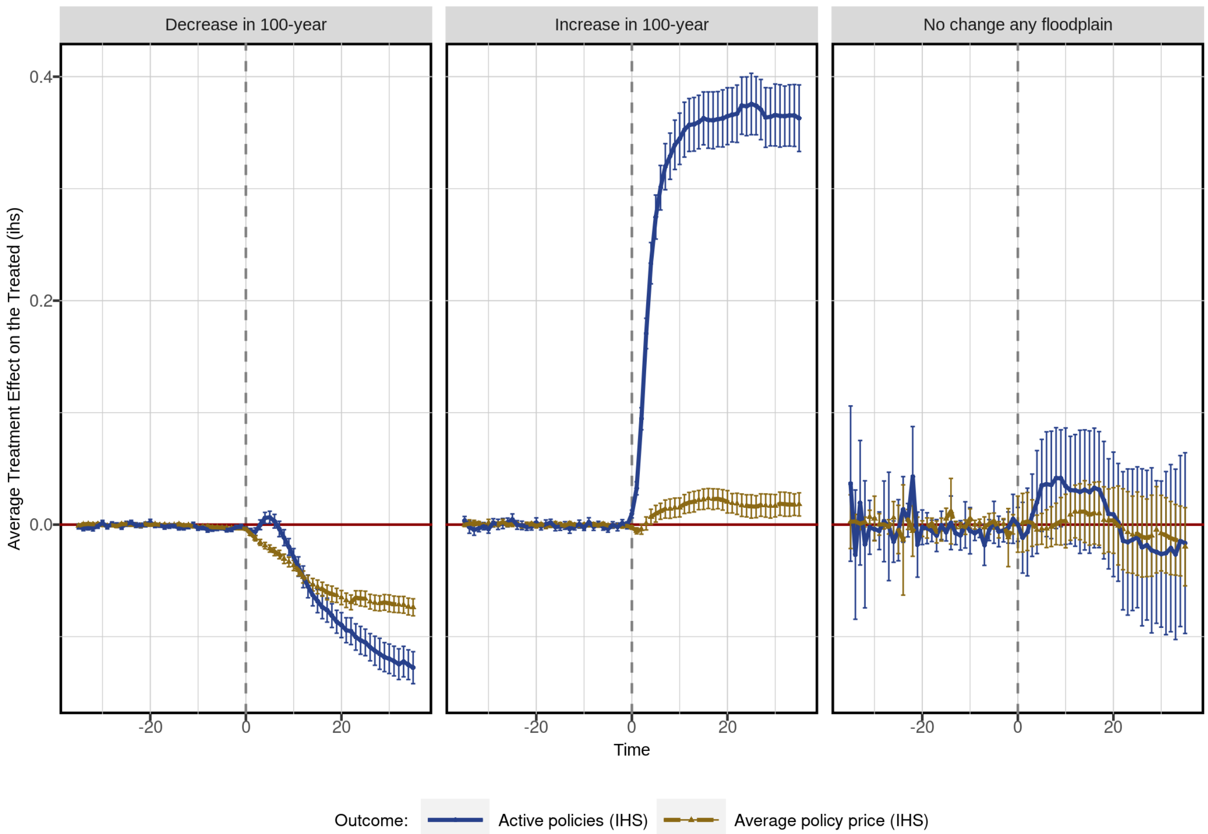
### 1.4.4 Event study estimates of the impacts of flood map updates

Figure 1.6 presents the Average Treatment Effect estimates by length of exposure, aggregated over all 114 cohorts following equation 1.4. The three panels estimate treatment effects separately based on whether the

[25] Additional analyses below further consider rezoning inside the 500-year floodplain

map update rezoned properties outside of the 100-year floodplain (first panel), inside (second panel), and where the map update did not cause any rezoning inside or outside of the 100-year and 500-year floodplains. The figure shows estimates for two different outcomes: the number of active flood insurance policies in the census tract (solid blue line) as well as the average price of the insurance policy (dashed gold line), both transformed using the inverse hyperbolic sine.

Figure 1.6: Aggregated event study estimates of the impacts of flood map updates on the demand for flood insurance and average policy price




Each facet represents average treatment effects for a different treated group, using treated census tracts where the flood map update increased or decreased the number of residential properties in the 100-year floodplain by more than 1% relative to the total number of residential properties in the census tract (first and second panel respectively). The third panel presents estimates for the treated census tracts that contained residential properties in the 100-year and 500-year floodplains, but where the map update did not change the number of properties located inside these floodplains. The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. The outcome variables are transformed using the Inverse Hyperbolic Sine (IHS). Error bars represent 95% confidence intervals using the multiplier bootstrap.

For all three treated samples and both outcome variables, we note the absence of pre-trends: the estimated average effects of receiving an updated flood map on demand and average prices is not significantly different from zero prior to the treatment date, increasing our confidence in the identifying assumption.

Post-treatment, however, effects are large and heterogeneous between samples and strongly suggest that being in the 100-year floodplain is a decisive driver of insurance take-up. Focusing first on the number of active policies, we note that among census tracts where the map modernization decreased the number of properties located inside the 100-year floodplain (first panel), the map update led to an average decline in the number of policies of about 10% after 2 years. In contrast, the map modernization caused an average increase of almost 40% in areas where the digital map rezoned properties inside the 100-year floodplain (second panel). In neighborhoods where the new flood maps did not rezone properties either inside or outside of the 100-year and 500-year floodplains (third panel), effects are small and not statistically different from zero.

In the appendix, I show in Figure B.1 that these effects (obtained aggregating models 1.3 over all treatment cohorts) also exist *within* each cohort, regardless of the treatment date. For most cohorts, the estimated impacts of map updates on the demand for insurance are positive for census tracts where the map updates rezoned properties inside the 100-year floodplain, and negative in census tracts where the map rezoned properties outside of the 100-year floodplain. Such patterns confirm the role the 100-year floodplain as a major driver of treatment effect heterogeneity.

Figure B.2 in the appendix shows these effects are not driven by new residential constructions: the estimated coefficients are extremely similar if we only retain insurance policies covering properties that were built prior to 2008 (the beginning of our insurance data sample) in the analysis. These results are also robust to changes in the transformation variable (log versus inverse hyperbolic spline), changes in the estimation sample (excluding all census tracts with less than 5 active insurance policies at all times), and controlling for the occurrence of FEMA disaster declarations (see Figure B.3 in the appendix, which combines these three robustness checks). Point estimates vary slightly based on the specific model considered, an effect driven by the exclusion of census tracts with few active insurance policies from the estimation procedure.

Figure 1.6 also shows the observed impacts on demand are not driven by changes in prices. The impacts of flood map updates on average prices follow the same patterns as the effects on the quantities, although the effects are more modest. Due to the aggregate nature of the data, interpretation of these effects is difficult. At the census tract level, average price is an equilibrium outcome: the effects measured here include both a direct floodplain effect (everything else equal, being in the 100-year floodplain increases the price of the insurance policy), as well as a composition effect (the pool of properties purchasing insurance might change with the new flood map). I highlight these effects in the next section by further separating policies that are purchased inside and outside the 100-year floodplain.

Overall, these results rule out any large effects of the pure digitization of paper-based maps, suggesting that access to information was not in itself a barrier to insurance take-up. On the other hand, the large impacts of the 100-year floodplain rezoning on insurance demand could be due to individuals revising their beliefs about flood risk following the new map, or because of the insurance mandate requirement inside the

100-year floodplain. The analysis below will partially differentiate between these two mechanisms.

### 1.4.5 Within-census tracts spillover effects of updated flood maps

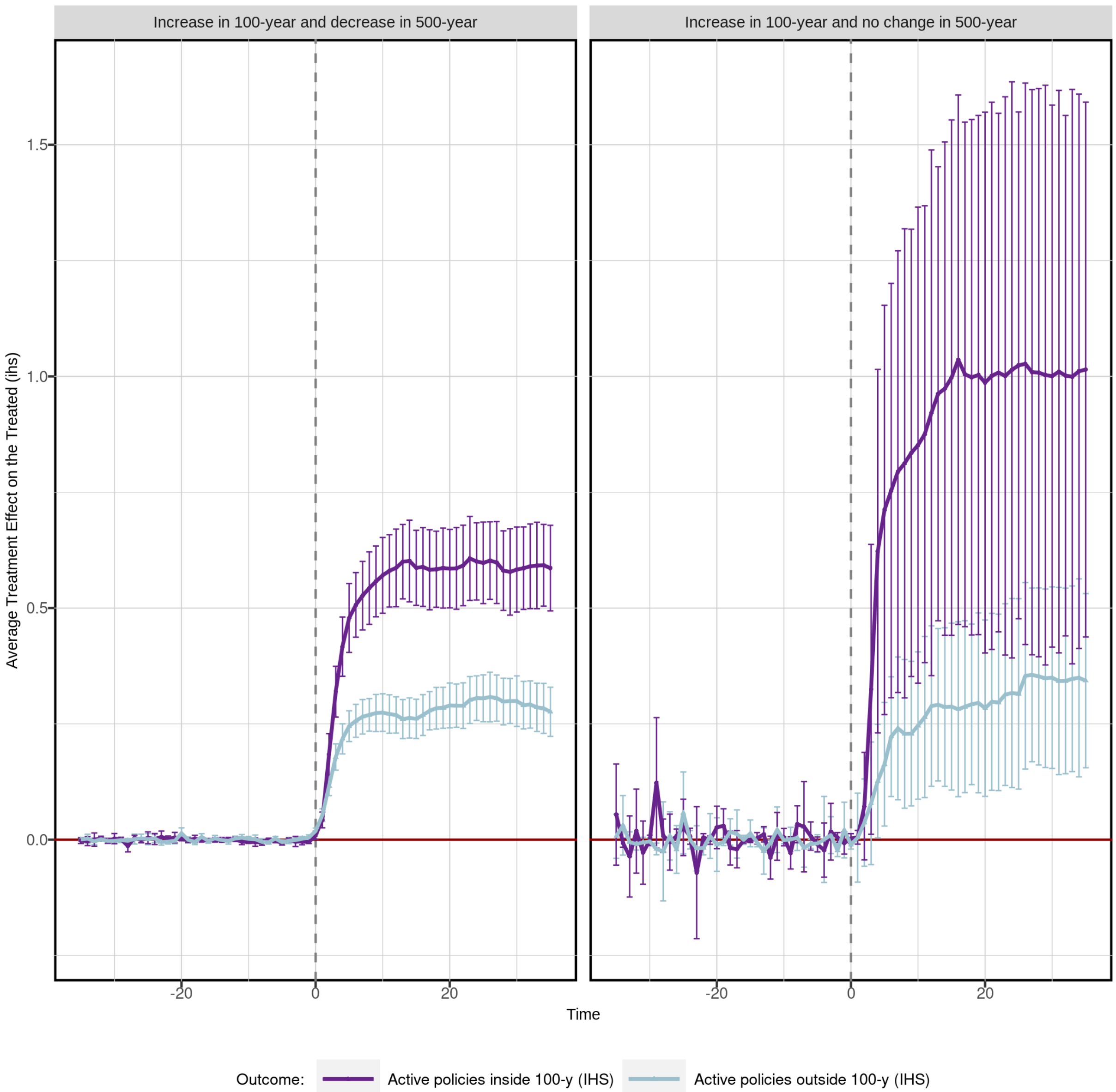
Flood maps present a discrete depiction of risk: the location of any property is either high-risk (100-year floodplain), low risk (500-year floodplain), or "minimal risk" (outside of the floodplains). Although flood maps provide some additional information within each zone, the categorical definition of flood risk has no underlying physical basis: absent specific structures, flood risk is in general continuous in space. This implies that homeowners might rationally form beliefs about their risk exposure beyond the discrete floodplain classification of their property. For instance, a homeowner could be located outside of the 100-year floodplain both before and after the map update, but might revise her beliefs about risk if the map update rezoned adjacent properties towards the 100-year floodplain.

The main challenge in testing for these "spillover effects" is that I cannot match insurance policies to geolocalized properties. However, the insurance policy data provide information on the floodplain in which the insured property is located *at the time of the policy purchase.* This last point bears emphasizing as it means that we can track, at any time, how many insurance policies are covering properties inside and outside of the 100-year floodplain, *within* census tract. However, we cannot interpret changes in the number of policies in a particular floodplain as changes in the insurance demand of properties that previously purchased a policy in this floodplain, since the *definition* of the floodplain itself is changing due to the map update.

Figure 1.7 presents heterogeneity-robust event-study estimates for the number of active insurance policies as outcome variables, this time differentiating the number of insurance policies *by floodplains within tracts* (recall that the definition of the floodplains changes at the time of treatment, since the floodplain classification comes from the insurance data). The models are only estimated for treated census tracts where the map updates rezoned properties inside the 100-year floodplain (these treated census tracts in Figure 1.7 were previously part of the first panel in Figure 1.6). Figure 1.7 further separates treatment samples based on the rezoning of properties in the 500-year floodplain: the left panel focuses on treated tracts where the map updates not only rezoned properties inside the 100-year floodplain but also rezoned properties outside the 500-year floodplain, while the right panel focuses on treated tracts where the map updates did not rezone any property either inside or outside of the 500-year floodplain.

As expected, in both cases I find that map updates increased demand for insurance from properties located inside the 100-year floodplain (purple lines). However, I find that these map updates also increased demand from properties located outside of the 100-year floodplain (light blue lines). The fact that this finding holds for census tracts where the map updates rezoned properties outside of the 500-year floodplain

Figure 1.7: Aggregated event study estimates of the impacts of flood map updates on the demand for flood insurance at the census tract and flood zone-level




Each facet represents average treatment effects for a different treated group, using treated census tracts where the flood map update increased the number of properties in the 100-year floodplain while simultaneously rezoning properties outside of the 500-year floodplain (left) or not rezoning any properties in the 500-year floodplain (right). The outcomes are measured at the *census tract / flood zone-level*, where the flood zone designations are derived from the insurance policy data. As such, they provide the number and cost of policies that are active within each zone at the time when the policy is purchased. The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. The outcome variables are transformed using the Inverse Hyperbolic Sine (IHS). Error bars represent 95% confidence intervals using the multiplier bootstrap.

(left) is all the more surprising: among these treated census tracts, multiple properties were rezoned from the 500-year floodplain to inside to 100-year floodplain and were likely carrying flood insurance – rezoning inside the 100-year floodplain would therefore mechanically cause the number of active insurance policies covering properties outside of the 100-year floodplain to *decrease* (recall that the number of insurance policies within each floodplain is derived from the insurance data itself and records the floodplain at the time of the purchase of the insurance policy). This means that the increase in insurance demand from properties located outside of the 100-year floodplain more than compensates this composition effect. Results are similar when focusing on census tracts where treatment did not rezone any properties either inside or outside of the 500-year floodplain (right), although effects are noisier in this second sample as the number of treated census tracts is smaller in this group.

Overall, these results demonstrate that rezoning properties inside the 100-year floodplain not only has a direct impact on insurance demand (by increasing the demand for insurance from properties located within the 100-year floodplain, as previously seen in the first panel of Figure 1.6), but also has a large indirect impact on properties located outside of the 100-year floodplain. While I cannot completely rule out the potential role played by the insurance mandate requirement inside the 100-year floodplain, these indirect or "spatial spillover" effects outside of the 100-year floodplain strongly suggest that a substantial portion of the impacts of updated flood maps on insurance demand is driven by changing beliefs about flood risks.

## 1.5 Distributional impacts of flood maps on household's demand

### 1.5.1 Limitations of the heterogeneity robust event-studies

The previous sections provide evidence of large average treatment effects of updating flood maps on the demand for flood insurance, including from properties that are not affected by the rezoning (spillover effect). However, the event-study specifications suffer from three important pitfalls.

First, one might be concerned that census tracts used as control units are inadequate. Although I use as controls the census tracts that later received a flood map with similar rezoning as the treatment group (either inside or outside of the 100-year and 500-year floodplains), one might still be worried that the control units fail to properly match the treated units. In particular, the Biggert-Waters Act of 2012 progressively phased out insurance subsidies for a subset of properties that were constructed prior to the community's first flood map (called "pre-FIRM properties") and are now located in the 100-year floodplain. The untreated potential outcomes of census tracts with higher shares of either pre-FIRM properties or properties located in the 100-year floodplain might therefore be on a different trend than those of census tracts with few pre-FIRM properties; failing to account for these differential trends could wrongly attribute the impacts of the BW Act to the flood map update.

Second, previous work found that the occurrence of a disaster is itself a strong determinant of flood insurance take-up (Gallagher, 2014). When estimating the impacts of flood map updates, one might therefore wish to account for the occurrence of disasters. Although the models estimated in Figure include dummy variables to control for previous FEMA disaster declarations, such controls might be inadequate if the response to disasters is highly non-linear.

Finally, although the average treatment effects estimated with the event-study above are robust to heterogeneity between cohorts, policy makers and floodplain managers are often interested in heterogeneity along other margins.[26] From a policy and floodplain management perspective, understanding *where* the policy effects are strongest is crucial.

One way to dive deeper into heterogeneous treatment effects and to construct better control groups – by accounting for the share of pre-FIRM properties, the share of properties zoned in the 100-year floodplain before the map update, and the potentially non-linear effects of disasters – would be to further split the samples of census tracts along these observable characteristics of interests. Unfortunately, this quickly becomes infeasible when using heterogeneity-robust estimators of the ATT such as the Callaway and Sant'Anna (2020) estimator, as the number of units in each $\{g, t, x\}$ cell quickly becomes too small – especially as the dimension of $x$ (the number of characteristics along which we wish to assess heterogeneity of the treatment effects) increases.[27] To escape this Catch-22, in the next section I estimate census tract-specific synthetic controls.

### 1.5.2 Estimating census tract-specific synthetic controls

In order to further investigate treatment effect heterogeneity and to alleviate the shortcomings of the event studies models presented in the previous section, I now turn to a refinement of the empirical approach of this paper: the estimation of local and clustered synthetic controls.

Synthetic controls were originally developed to estimate treatment effects when only a small number of distinct units are available, such as states or countries, and when only one unit receives treatment (Abadie

[26] Figure B.1 presents cohort-specific treatment effects, but within each cohort some neighborhoods might respond strongly to treatment while other may respond little – either due to variation in policy implementation (rezoning intensity) or due to variation in unit-specific characteristic (initial share of properties in the 100-year floodplain, income, recent occurrence of a flood, etc). This critic could be applied to any estimator which targets some average treatment effect, but is especially relevant in our case, as the distribution of the outcome variable is extremely skewed: the demand for insurance is low in most census tracts, and very large in a few hundred census tracts.

[27] A "standard" approach to assess heterogeneity would be to run two-way fixed effect regressions and interact treatment with census tract-specific covariates. As discussed above, this would deliver biased and inconsistent treatment effect estimates due to both between-cohort heterogeneity in the treatment effect and treatment effect dynamics (which was the entire rationale for using the Callaway and Sant'Anna (2020) estimator in the first place).

and Gardeazabal, 2003; Abadie et al., 2010). Because directly comparing the outcome of the treated unit to the outcomes of the control units generally provide a poor estimation of the treatment effect, a "synthetic control" is first constructed by taking a weighted average of the different control units, where the weights are chosen so that the evolution of the synthetic control's outcome pre-treatment matches as closely as possible the evolution of the outcome of the treated unit. The construction of the synthetic control can also account for other covariates or multiple outcome variables, such as in Robbins et al. (2017).

I estimate unit-specific synthetic controls for each census tract observed at least 12 months pre-treatment and 24 months post-treatment. As discussed in Abadie (2020), multiple applications of the synthetic control method only consider matching on the outcome variable pre-treatment. Although matching on the pre-treatment value of other covariates can only decrease pre-treatment fit on the outcome variable, such matching can improve the credibility of synthetic control estimates.[28] I generate synthetic controls by matching on pre-treatment values of the outcome variable (the number of active flood insurance policies in the census tract), as well as pre-treatment values of (i) the share of pre-FIRM properties and (ii) the share of policies covering properties in the 100-year floodplain. Since census tracts with similar shares of pre-FIRM properties and properties covered by insurance in the 100-year floodplain are likely to similarly respond to the Biggert-Waters Act and other idiosyncratic time shocks, matching on these two covariates generate more credible synthetic controls.

An important choice that remains before the estimation concerns the definition of the "donor pool", i.e., the group of units that do not receive treatment and which are considered when estimating the synthetic control weights. In a typical synthetic control application, all untreated units are part of the donor pool. However, in our case, flooding and other spatially correlated disaster events cause large increases in the demand for flood insurance not only in areas directly impacted by flooding, but also in neighboring areas that are located in the same television-network market (Gallagher, 2014). During our observation window, flooding occurs in multiple states. Not accounting for such flooding events and their spatial spillover effects could generate spurious synthetic control estimates, since the potential outcomes of units experiencing a flood are likely to diverge. In order to mitigate this issue, I estimate *spatially clustered* synthetic controls: for each census tract that receives treatment, the "donor pool" contains units that do not receive treatment within our observation period and that are located within the *same* FEMA region.[29]

[28] Abadie (2020) emphasizes this aspect: "Part of the literature on synthetic controls emphasizes estimators that depend only on pre-intervention outcomes and ignore the information of other predictors [...]. This reliance on pre-intervention outcomes only, while adopted in many cases for technical or expositional convenience, may create the mistaken impression that other predictors play a minor role in synthetic control estimators."

[29] One possibility is to form the donor pool using units that are not-yet-treated, instead of units that are never-treated. This is appealing, since not-yet-treated units are more likely to resemble already-treated units. However, this makes it hard to interpret changes in the treatment effects over time, since some of these changes are driven by changes in the composition of the donor pool.

Finally, to improve the quality of pre-treatment fit, I use the augmented synthetic control method presented by Ben-Michael et al. (2020). This approach uses an explicit outcome model to estimate how much bias in the treatment effect can be expected from imperfect pre-treatment fit, and then uses a correction to remove bias in the original synthetic control estimates. Formally, the estimator proceeds in two steps. First, we solve the "standard" synthetic control program:

$$\begin{aligned}
&\min_{\gamma} \quad \theta_x \left\| \boldsymbol{X}_1 - \boldsymbol{X}_0' \gamma \right\|_2^2 + \theta_z \left\| \boldsymbol{Z}_1 - \boldsymbol{Z}_0 \cdot \gamma \right\|_2^2 + \zeta \sum_{W_i=0} f\left(\gamma_i\right) \\
&\text{subject to} \sum_{W_i=0} \gamma_i = 1 \\
&\qquad \gamma_i \geq 0 \quad i : W_i = 0
\end{aligned} \tag{1.5}$$

where the goal is to find the vector or weights $\gamma$ in the simplex

$\Delta^{N_0} = \left\{ \gamma \in \mathbb{R}^{N_0} \mid \gamma_i \geq 0 \ \forall i, \ \sum_i \gamma_i = 1 \right\}$ that minimizes the synthetic control objective function. This function is made of three parts: (i) the L2-norm (Euclidean distance) between the pre-treatment outcome of the treated census tract $\boldsymbol{X}_1$ and the control census tracts $\boldsymbol{X}_0$, (ii) the L2-norm between the pre-treatment covariates of the treated census tract $\boldsymbol{Z}_1$ and the control census tracts $\boldsymbol{Z}_0$, and (iii) a term that penalizes the dispersion of the weights assigned to control units (those with $W_i = 0$), for some function $f$ and a positive hyperparameter $\zeta$.[30] The weights $\theta_x$ and $\theta_z$ govern the relative importance of the deviations between lagged outcomes and covariates in the minimization program.[31]

In a second step, we "augment" the synthetic control to estimate the (counterfactual) potential outcome of the treated unit:

$$\begin{aligned}
\hat{Y}_{1T}^{\text{aug}}(0) &= \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} Y_{iT} + \left( \hat{m}_{1T} - \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \hat{m}_{iT} \right) \\
&= \hat{m}_{1T} + \sum_{W_i=0} \hat{\gamma}_i^{\text{scm}} \left( Y_{iT} - \hat{m}_{iT} \right)
\end{aligned} \tag{1.6}$$

where $\hat{\gamma}_i^{\text{scm}}$ are the solutions to the program in equation 1.5, $Y_{iT}$ are the post-treatment outcomes, and $\hat{m}_{iT}$ is an estimator of the post-treatment control potential outcome for unit $i$. In the standard synthetic control case, $\hat{m}_{iT}$ is just a constant. I follow Ben-Michael et al. (2020) and use a ridge regression for the outcome model.[32]

[30] The initial applications of the synthetic control methods did not include this penalty term – it is discussed in footnote 10 of Abadie et al. (2015) as a way to select weights when the minimization of the other parts of the objective function has multiple solutions. Different choices of penalty functions exist; see for instance Doudchenko and Imbens (2016) for a discussion. In this paper I implement the Ridge-Augmented synthetic control approach, for which Ben-Michael et al. (2020) showed that the penalty term replaces the simplex constraints with the form $f\left(\gamma_i\right) = \left(\gamma_i - \hat{\gamma}_i^{\text{scm}}\right)^2$ (deviations from the standard synthetic control weights are penalized).

[31] Following Ben-Michael et al. (2020), in my preferred approach I use $\theta_x = \theta_z = 1$.

[32] This choice of estimator has attractive properties, notably an improvement in pre-treatment fit relative to the standard synthetic control model, and a reduction in estimation error under linear and latent-factor data generating processes.

(Augmented) synthetic controls allow for the flexible estimation of treatment effect heterogeneity and can uncover whether differences in average treatment effects between groups are only driven by a small number of units within each group. They also greatly mitigate the shortcoming of the event studies presented above. First, treatment effects are estimated separately for each census tract that receives a new flood map during our observation window. Under the standard Stable Unit Treatment Value Assumption, the staggered nature of the treatment does not contaminate the estimated treatment effect.[33]

Synthetic controls also make it possible to escape the "micronumerosity" problem. Section 1.5.1 highlighted how the staggered nature of the treatment combined with treatment effect heterogeneity required the estimation of average treatment effects in small cells defined by treatment timing (for instance, we estimate a distinct average treatment effect among the group of census tracts treated in August 2009). Further splitting these cells according to the value of other neighborhood-level characteristics (such as the number of insured properties in the 100-year floodplain or the share of pre-FIRM properties) quickly lead to numbers of observations that are too small to estimate the (cell-level) average-treatment effects. In contrast, the synthetic controls allow us to estimate treatment effects in the extreme case where each cell contains only one unit.

Moving from the estimation of average treatment effects to neighborhood-specific treatment effects is not without costs. One major drawback is that synthetic controls can perform poorly when the treatment effect is small relative to the noise in the outcome variable (Abadie, 2020). Unlike with difference-in-differences estimators, the statistical power of synthetic control estimators does not increase with sample size. In our context, this means that we are more likely to credibly estimate treatment effects where flood map updates lead to large changes in insurance take-up, and for census tracts where there is a large number of active insurance policies.[34] I therefore estimate synthetic controls in census tracts that have at least 20 active policies at all times.[35]

Finally, quantifying the uncertainty around synthetic control estimates is still an active area of research. To the best of my knowledge, there is no agreement on how to account for multiple hypothesis testing when evaluating the uncertainty around the treatment effects of thousands of synthetic controls. Despite these shortcomings, I use the recently developed jackknife+ procedure from Barber et al. (2020) to construct confidence intervals.

[33] Note that the spillover effects uncovered in section 1.4.5 are not inconsistent with the Stable Unit Treatment Value Assumption, as these spillovers are found within each census tract, and not between different tracts.

[34] Recent extensions of synthetic control estimators consider how to estimate the Average Treatment Effect on the Treated when several units receive treatment at different times (Xu, 2017; Robbins et al., 2017; Ben-Michael et al., 2019).

[35] Due to the spatial concentration of the demand for flood insurance, excluding census tracts with less than 20 active insurance policies at all times lead to a negligible reduction in the aggregate number of active policies considered for the synthetic control estimation.

### 1.5.3 Synthetic controls estimates

This section presents the (ridge-augmented) synthetic control estimates for each of the 4870 census tracts that have at least 20 active policies at all times, for which I can identify changes in the 100-year floodplain, and that received a digital flood map within the period 2009-2017 using the procedure described in the previous section.

To assess pre-treatment fit of the synthetic control estimates, I compute the normalized root mean squared error (NRMSE) during the pre-treatment period: for 77 % of the estimated synthetic controls, the NRMSE is less than 0.025, meaning that the synthetic control and the treated census tract deviate by less than 2.5% on the number of policies prior to treatment in most cases.[36] A histogram of the NRMSE is presented in Figure C.1 in the Appendix. In the following sections, I restrict the analysis to census tracts for which I estimated a synthetic control that provides a close match to the pre-treatment outcome, defined as a NRMSE lower than 0.05.[37] In our application, the smallest donor pool has 55 units, while the largest donor pool has 839 units. Figure C.2 shows that the pre-treatment fit is not correlated with the number of units in the donor pools.

Figure 1.8 presents the dynamic treatment effects of the impact of flood map updates on insurance demand estimated through synthetic controls for six selected census tracts with 90% confidence intervals (shaded bands) obtained through jackknife+ (Barber et al., 2020). Pre-treatment fit is good for the first five synthetic controls, while the sixth synthetic control estimates present a NRMSE greater than 5%, and is excluded from the remainder of the analysis. Four out of the six treatment effect estimates present an immediate and sustained response to treatment, ranging from -300 flood insurance policies in the second panel to +70 insurance policies in the fifth panel.

While medium and long-term treatment effects are most relevant from a public policy perspective, estimating long-term treatment effects require stronger assumptions and more data. In order to strike a balance between these, for the remainder of this section I focus on the estimated treatment effects at 24 months post-treatment.[38]

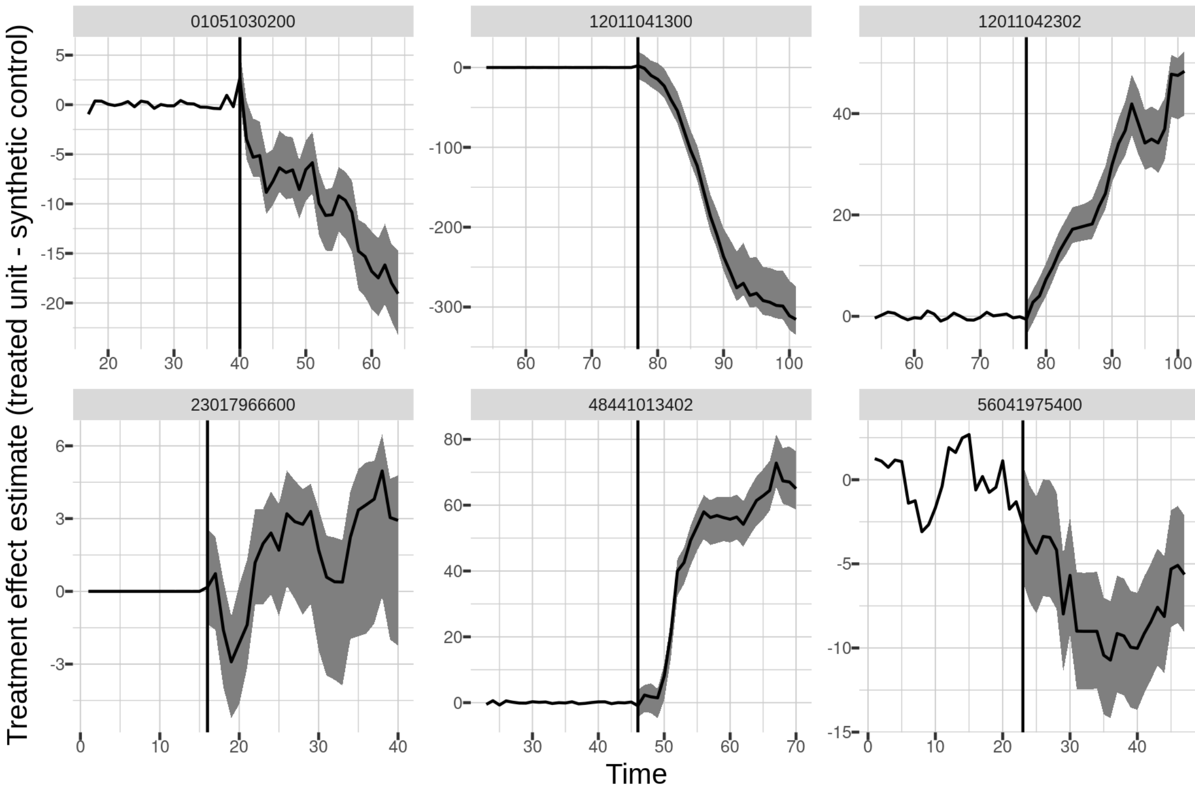
Figure 1.9 depicts all the synthetic control estimates for census tracts with at least 20 active insurance policies at all times. Each dot is therefore a treatment effect estimated for one specific census tract, with the point estimate taken 24 month post-treatment. The treatment effects are expressed relative to value of the synthetic control, and are plotted against the log of the average number of active policies during the pre-

[36] It is smaller than .01 in 64 % of cases.

[37] Most of the deviations between the treated units and their estimated synthetic control outcomes pre-treatment arise from the inclusion of covariates in the matching procedure: when omitting covariates, the weighting procedure generates weights that lead to NRMSE lower than 0.01 in 92% of cases.

[38] Additional results show that effects are similar when focusing on treatment effects after 12 months or 30 months instead.

Figure 1.8: Synthetic control estimates of the impacts of flood map updates on the demand for flood insurance in six census tracts

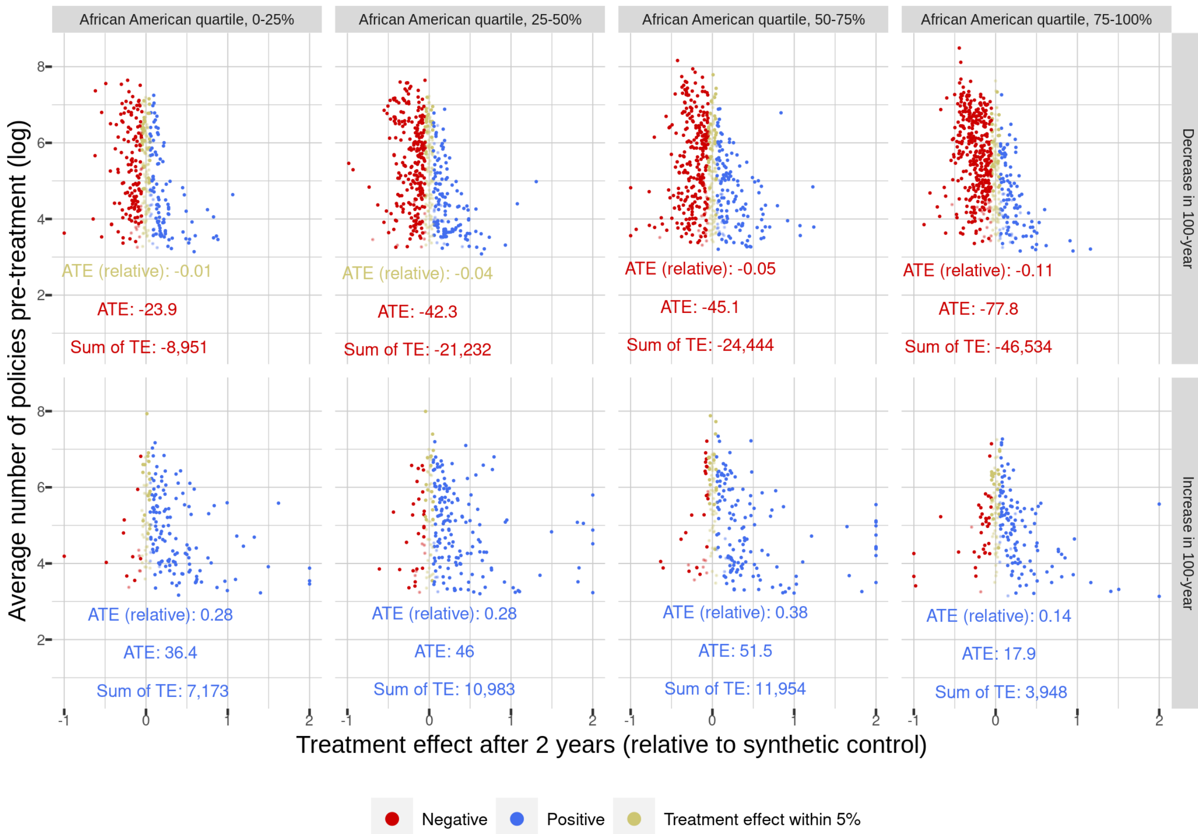



Dynamic treatment effects estimates of the impacts of flood map updates on the demand for flood insurance in six census tracts, estimated by synthetic controls augmented with ridge regression. The solid black line shows the difference between the census tract outcome (number of active policies) and the synthetic control constructed for this tract. The vertical bar denotes the end of the pre-treatment optimization period, and the grey ribbon depicts 90% jackknife+ confidence intervals.

treatment period. This representation allows to examine whether large relative treatment effects (x-axis) are concentrated in census tracts with a low or high demand for insurance prior to treatment (y-axis). Large dots depict treatment effects that are significantly different from 0 at the 10% level using the jackknife+ inference procedure. The spatial location of these treatment effects is presented in Figure C.3 in the appendix.

In order to investigate treatment effect heterogeneity and the distributional impacts of map updates, the two rows of panels depicts treatment effects based on the rezoning of residential properties inside or outside of the 100-year floodplain (top and bottom row respectively). The four columns separate the census tract-specific treatment effects by quartile of the African American population share, with the first column depicting census tracts with the highest share of White individuals. Finally, within each panel I compute the relative Average Treatment Effect (average of the synthetic control estimates divided by the number of active policies pre-treatment within each tract), the (overall) Average Treatment Effect (average of the synthetic control estimates), and the aggregate treatment effect (sum of the synthetic control estimates).

Figure 1.9: Estimated synthetic control treatment effects of the impact of the map update on demand for insurance, by quartiles of neighborhoods' share of African Americans.




Each dot represents a census tract-specific treatment effect estimate of the impact of the flood map update on the demand for flood insurance after 24 months. Red and blue dots show negative and positive treatment effects, respectively. The facets separate census tracts where the flood map update increased or decreased the number of properties zoned inside the 100-year floodplain by more than 1% (first and second rows respectively). Census tracts are further separated by quartiles of African American population in the census tract (columns). Treatment effects are estimated using synthetic controls augmented by ridge regression. For each treated unit, the donor pool comprises never-treated census tracts within the same FEMA region.

We first note that there is substantial heterogeneity in the estimates: positive and negative treatment effects occur in all six panels and for all values of insurance demand pre-treatment. However, there is a clear excess of large, negative treatment effects among the census tracts where the map updates rezoned properties outside of the 100-year floodplain (first row). In these census tracts, the map updates caused an aggregate decline in insurance take-up of about 100,000 policies (shown on the panels as the "Sum of TE"). This decline is driven by treatment effects in census tracts with a large number of active policies pre-treatment, as shown by the small relative Average Treatment Effects (from -1% in the first panel to -11% in the fourth panel) and the comparatively large Average Treatment Effect values (from -23.9 to -77.8 policies).

In contrast, among census tracts that rezoned properties inside the the 100-year floodplain (second row of Figure 1.9), the aggregate effect are positive and the average treatment effects relative to the synthetic controls is also large, varying from +14% to +38%, suggesting that the positive effect of the map update on demand occurs for census tracts with both low and high pre-treatment demand for insurance.

Figure 1.10 shows the spatial distribution of the estimated treatment effects, with a focus on the area where hurricanes Laura and Ida made landfall in August 2020 and 2021, respectively. We note that negative and positive treatment effects are concentrated in clusters of effects having the same sign. Indeed, as shown in Figure 1.4, the map updates tended to rezone properties in a similar way for neighborhoods located close to one another. Furthermore, as shown above, negative treatment effects are strongly associated with rezoning of properties outside of the 100-year floodplain – these two facts generate the spatial clustering of treatment effects observed in Figure 1.10. However, Figure 1.10 also depict that neighborhoods close to one another sometimes appear to respond very differently to the update of flood maps (for instance in New Orleans).

Overall, these synthetic control estimates confirm the results obtained through heterogeneity-robust difference-in-differences models: rezoning properties inside the 100-year floodplain causes an increase in the demand for insurance on average, while rezoning outside of the 100-year floodplain causes a decline in the demand for insurance. They also highlight a substantial underlying heterogeneity, which the rest of this section explores in more details.

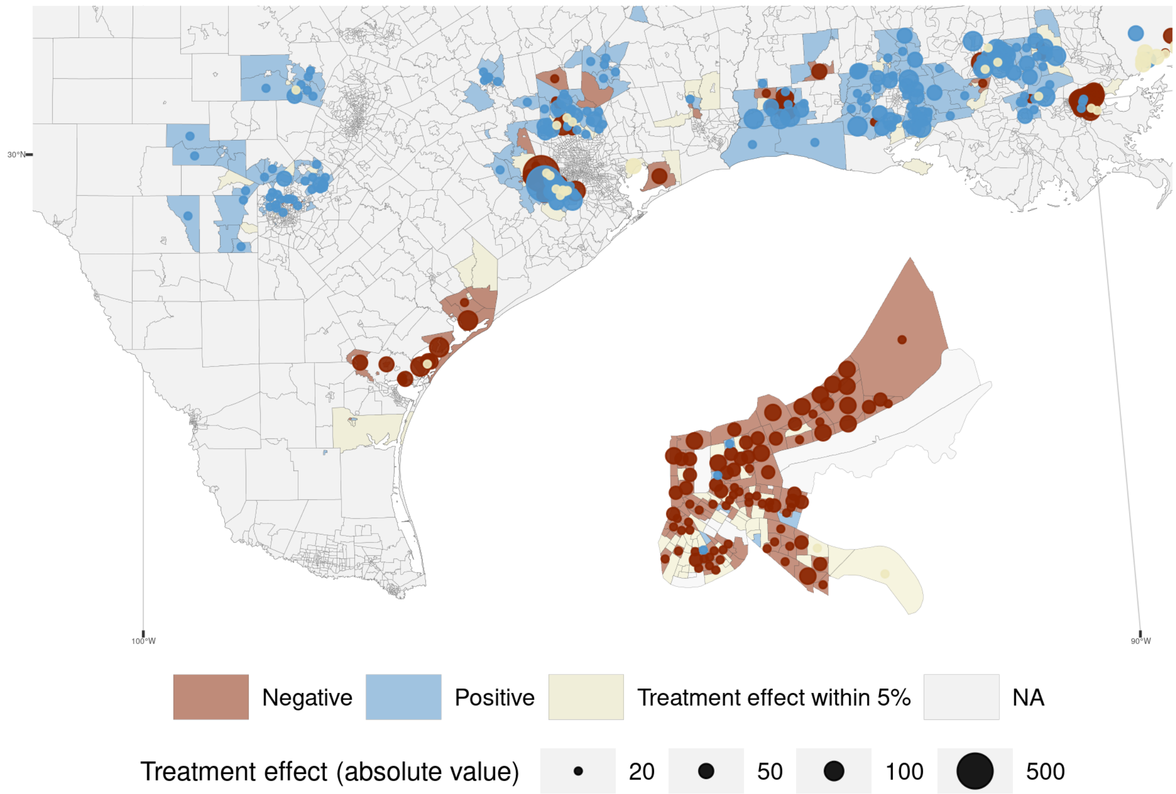
### 1.5.4 Distributional treatment effects and mechanisms

Figure 1.9 reveals that while rezoning outside of the 100-year floodplain caused substantial declines in the demand for flood insurance, these declines are unequally distributed between census tracts. Most of the large, negative treatment effects occur in the top quartile of the census tract African American population share (fourth column): negative treatment effects in these neighborhoods account for more than half of the map-induced aggregate decline in the demand flood insurance policies.

Such unequal impacts between "White" and "Black" neighborhoods are surprising. The disaster insurance literature and the stylized model of section 1.2.2 propose three potential mechanisms that could rationalize the observed heterogeneity: (i) differences in income levels, (ii) differences in treatment implementation, and (iii) differences in beliefs about flood risk.

Black neighborhoods tend to be poorer on average. If individuals have decreasing absolute risk aversion, then households in Black neighborhoods might be more sensitive to rezoning inside and outside of the 100-year floodplain. On the other hand, map updates varied in *treatment intensity* between neighborhoods: as seen in section 1.4.1 some maps rezoned many properties inside and outside of the 100-year floodplain, while other map updates were pure digitizations. If map updates were more likely to rezone properties inside the 100-year floodplain in White neighborhoods and outside of the 100-year floodplain in Black neighborhoods, this

Figure 1.10: Synthetic control estimates in the area recently impacted by hurricanes Laura and Ida




Each dot represents a census tract-specific treatment effect estimate of the impact of the flood map update on the demand for flood insurance. Red and blue dots show negative and positive treatment effects, respectively. The size of the dot represents the absolute value of the treatment effect.

would lead to unequal impacts of flood map updates between neighborhoods, even if income did not impact the demand for flood insurance. Finally, households beliefs about flood risk might mediate the insurance demand response to new maps. While beliefs cannot be observed, they respond to previous exposure to natural disasters (Gallagher, 2014). If Black neighborhoods were less likely to be hit by a natural disaster prior to their flood map updates, this could dampen the response to "official" flood risk information and also explain the observed distributional effects.

To test these three mechanisms, I run the following regression:

$$\widehat{TreatmentEffect}_i = \alpha + \beta Rezoning100year_i \cdot IncomeQ_i \cdot DisasterDeclaration_i + \epsilon_i \tag{1.7}$$

where $\widehat{TreatmentEffect}_i$ is the synthetic control estimate of the effect of the map update on the demand for flood insurance in census tract $i$, $Rezoning100year_i$ indicates the number of residential properties rezoned inside the 100-year floodplain with the map update (relative to the number of residential properties in the census tract), $IncomeQ_i$ indicates to which income quartile the census tract belongs to, while

$DisasterDeclaration_i$ keeps track of whether there was an official FEMA disaster declaration in census tract $i$ within the two years prior to treatment.[39]

Figure 1.11 depicts the marginal effects of regression 1.7 along with the treatment effect estimates. As in Figure 1.9 above, each dot represents a census tract-specific synthetic control estimates at 24 months post treatment. The x-axis depicts the relative change in the number of properties in the 100-year floodplain due to the map update (treatment intensity, with rezoning outside of the 100-year floodplain counted negatively), while the y-axis depicts magnitude of the relative treatment effects. The left panel focuses on census tracts where the map update occurred within two-years of a FEMA disaster declaration, while census tracts in the right panel did not experience a disaster declaration before treatment.[40] The different colors represent census tracts in different income quartiles. The lines represents the marginal effects of increasing the share of properties rezoned in the 100-year floodplain for each of the income quartile and disaster declaration groups.

We first note a strong relationship between treatment intensity and the magnitude of the treatment effect: the more properties rezoned inside the 100-year floodplain, the larger the treatment effect. The overall effect of rezoning inside the 100-year floodplain in regression 1.7 is 0.19 , while the adjusted-$R^2$ of regression 1.7 is 65 %. Put differently, the three variables in regression 1.7 explain two thirds of the variation in the synthetic control estimates. The largest negative treatment effects are concentrated in neighborhoods where the map updates rezoned a large share of properties outside of the 100-year floodplain – explanation (ii) above appears to substantially contribute to the observed distributional impacts between Black and White neighborhoods.

On the other hand, Figure 1.11 reveals that income is unlikely to drive the distributional effects of the map updates. The marginal effects of rezoning properties inside the 100-year floodplain *increases* (in absolute value) with income, suggesting that higher-income households are *more* sensitive to the rezoning of properties inside and outside of the 100-year floodplain. This pattern is inconsistent with decreasing absolute risk aversion, but could be rationalized by a different behavioral model or differential access to information between poorer and wealthier neighborhoods.[41]

Finally, Figure 1.11 shows that the relationship between rezoning of properties, income, and census tract specific treatment effects is strikingly similar between neighborhoods which experienced a FEMA disaster declaration in the two-years prior to the flood map update.

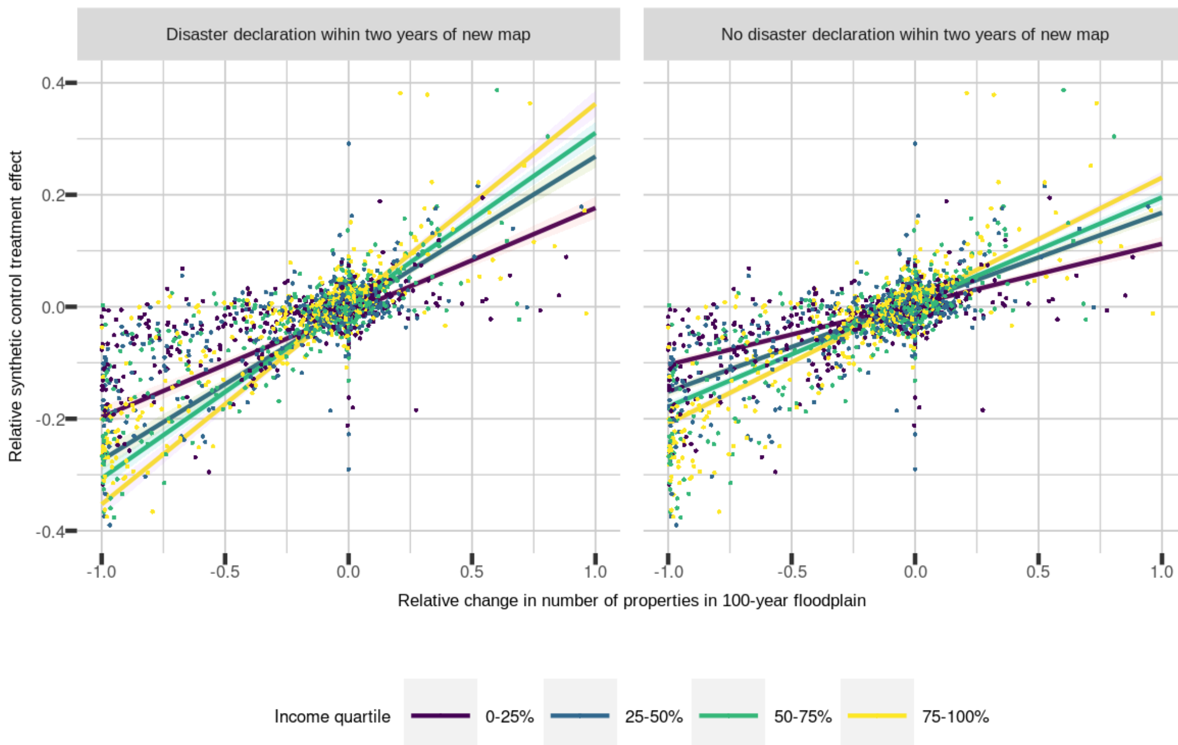
Overall, these results highlight that the distributional impacts of the map updates on the demand for

[39] Note that because the left-hand side of regression 1.7 is estimated using synthetic controls, the standard errors need to be interpreted with caution. To the best of my knowledge, this "second stage" regression (where the first stage is the construction of the synthetic control estimates) has not been considered previously and its theoretical properties have not yet been derived. This is left for future developments.

[40] Results are similar if we use different cut-off thresholds for the number of years prior to a disaster declaration. I used two-years for convenience, as roughly half of census tracts fall on each side of this cut-off.

[41] Outreach efforts regarding the updated flood maps may vary between neighborhoods, but despite multiple requests I was not able to obtain data on the magnitude of outreach efforts in different communities.

Figure 1.11: Second-stage regression, marginal effects by income and disaster declaration exposure




Marginal effects of a change in the number of properties rezoned inside the 100-year floodplain on the synthetic control treatment effect estimates, following regression 1.7. The marginal effects are estimated separately per income quartile and disaster declaration group. Each dot represents a census tract-specific treatment effect estimate of the impact of the flood map update on the demand for flood insurance after 24 months using synthetic controls augmented by ridge regression. For each treated unit, the donor pool comprises never-treated census tracts within the same FEMA region.

disaster insurance are largely due to heterogeneity in the magnitude of treatment (i.e., the rezoning of properties inside and outside of the 100-year floodplain). Map updates were more likely to rezone properties outside of the 100-year floodplain in Back neighborhoods. Because the magnitude of rezoning is associated with larger treatment effects, this caused substantial declines in the demand for flood insurance. Conditional on a given treatment magnitude, wealthier neighborhoods appear more sensitive to the map updates, but this effect is not enough to offset the impacts of greater rezoning inside the 100-year floodplain.

### 1.5.5 Discrepancies between official flood maps and independent flood risk estimation

Up until this point, we focused on the rezoning of the 100-year floodplain as it appears on the official flood maps produced by FEMA. For the past sixty years, these maps were the main source of flood risk information available to the public, if not the only source.[42] This recently changed. In 2020, a consortium of scientists, technologists, and communicators called the First Street Foundation (FSF) started producing and disseminating flood risk estimates for the entire contiguous United States, at a very fine level of spatial precision (3 meters). The FSF Flood Model combines historical data on past flooding events, synthetic coastal cyclones simulations, observed sea level rise and tidal levels with region-specific hydraulic models that account for human adaptation infrastructures to estimate the risk of flooding at the property-level, and is considered to provide the best residential flood risk estimates in the US (First Street Foundation, 2020; Armal et al., 2020). In this section, I compare flood risk estimates using the (FEMA) flood maps and the FSF Flood Model to assess whether the official map updates were "correct" – assuming that the truth is provided by the FSF Flood Model. I then use these estimates to assess whether the households' demand for insurance responds to the "correctness" of the map updates.

Table 1.3 summarizes the number of properties located in the 100-year floodplain as defined on the Q3 flood maps, the digital flood maps (as of December 2019), and the FSF flood model. More than 14.8 million properties are located in the 100-year floodplain based on the FSF flood model, while less than half are located in the 100-year floodplain as shown on the official flood maps in 2019. This discrepancy was highlighted in previous work that found official flood maps tended to under-estimate the extent of the 100-year floodplain, in particular due to systematic omissions of pluvial flood risk Wing et al. (2018). In addition, Table 1.3 shows that although the Q3 flood maps only covered about two thirds of residential properties nation-wide (87 million properties), it depicted 7.2 million properties in the 100-year floodplain, or 500,000 *more* than are depicted today in the digital FEMA flood maps – despite the fact that current maps cover 30 million more properties than the Q3 maps did. This highlights an important point: the deficit of properties in the official 100-year floodplains (as compared to the FSF flood model) is *not* driven by outdated flood maps, since the Q3 flood maps are a digitization of paper-based maps that were available in 2005. As the next figure shows, the deficit of properties mapped in the 100-year floodplain is driven by the updating process itself.

Figure 1.12 focuses on areas that were mapped in both the Q3 flood product (in 2005) as well as in 2019 (digital flood maps) to decompose how modernization of flood maps impacted the number of properties

[42] Several insurance and re-insurance companies developed their own models of flood risk, but these were not publicly available.

Table 1.3: Number of residential properties located in the 100-year floodplain in different flood risk mapping products

| Flood zone | Q3 | NFHL 2019 | First Street Model |
|---|---|---|---|
| Inside 100-year floodplain | 7,200,000 | 6,700,000 | 14,800,000 |
| Outside 100-year floodplain | 80,000,000 | 109,600,000 | 117,500,000 |
| Not mapped | 45,100,000 | 16,000,000 | 0 |

zoned inside of the 100-year floodplain. The "correct" floodplain status is assumed to be provided by the FSF flood model. There are three important conclusions. First, while we previously saw that map updates tended to remove properties from the 100-year floodplain, comparisons with the FSF flood model indicates that a substantial share of the removals appear to be warranted: among the 2.4 million properties that were rezoned outside of the 100-year floodplain during the map updates, 1.7 million properties were correctly removed. In contrast, half of the 1.1 million properties that were zoned inside the 100-year floodplain during the flood map update appear to have been incorrectly added. Finally, more than 5.2 million properties were mapped outside of the 100-year floodplain in 2005 and should have been rezoned inside the 100-year floodplain, yet were left out during the map update.

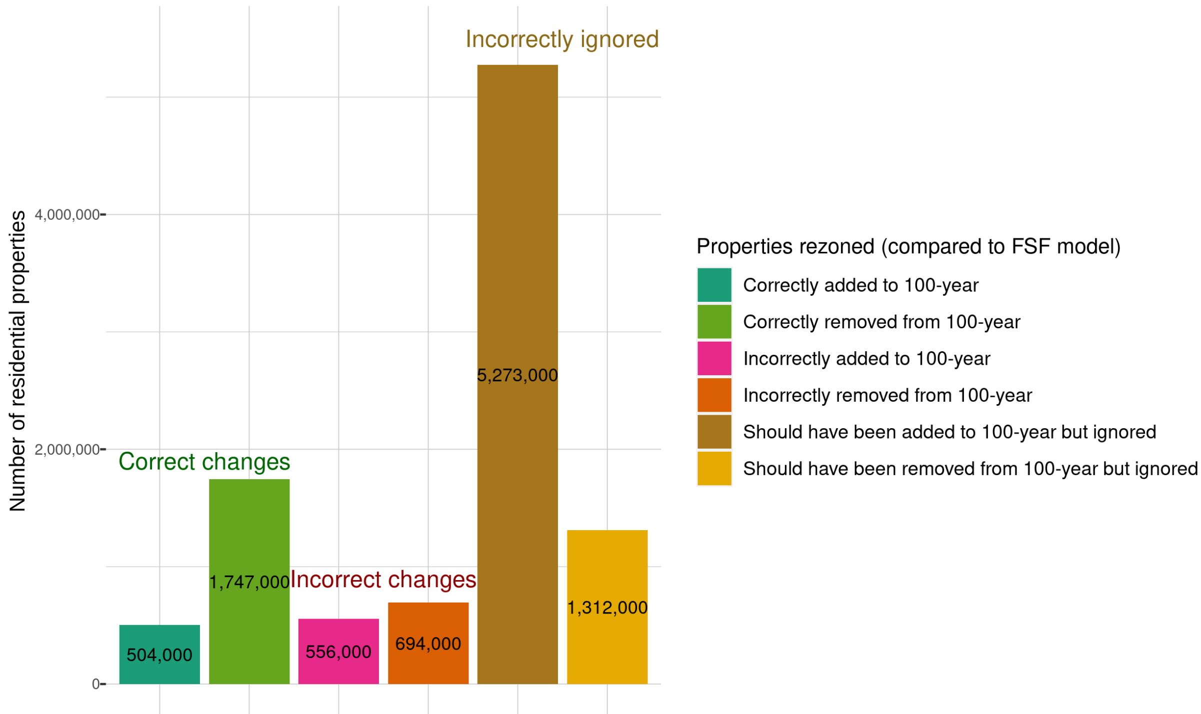
These descriptive results reveal that the modernization of flood maps primarily failed by *omission*: while two thirds of the removals from the 100-year floodplain seem to have been warranted, more than five million properties were incorrectly "ignored" during the modernization process when they should have been added to the 100-year floodplain (and 1.3 million properties should have been removed from the 100-year floodplain). Assuming that the FSF Flood Model is closer to the "true flood risk" than the FEMA flood maps, this suggests current official map updates are insufficient to convey flood risk information for individuals who live in the 100-year floodplain. Outdated flood maps are an issue, but updating flood maps without significantly altering incorrect floodplain boundaries will not solve it.

### 1.5.6 Impact of information correctness

It is a priori possible that households hold private information about their *true* flood risk exposure. Under this scenario, households could selectively respond to flood map updates that are "correct", and not substantially change their demand for insurance following an incorrect flood map update. Such differentiated response to flood map updates would mitigate the welfare costs of providing incorrect information, and could explain part of the heterogeneity in the treatment effect estimates highlighted above. Unfortunately, the analysis below shows that this is not the case.

Figure 1.13 presents the synthetic control estimates at 24-months, plotted against the relative change in

Figure 1.12: Changes in the number of residential properties in the FEMA 100-year floodplain between 2005 and 2019, compared to the FSF flood model



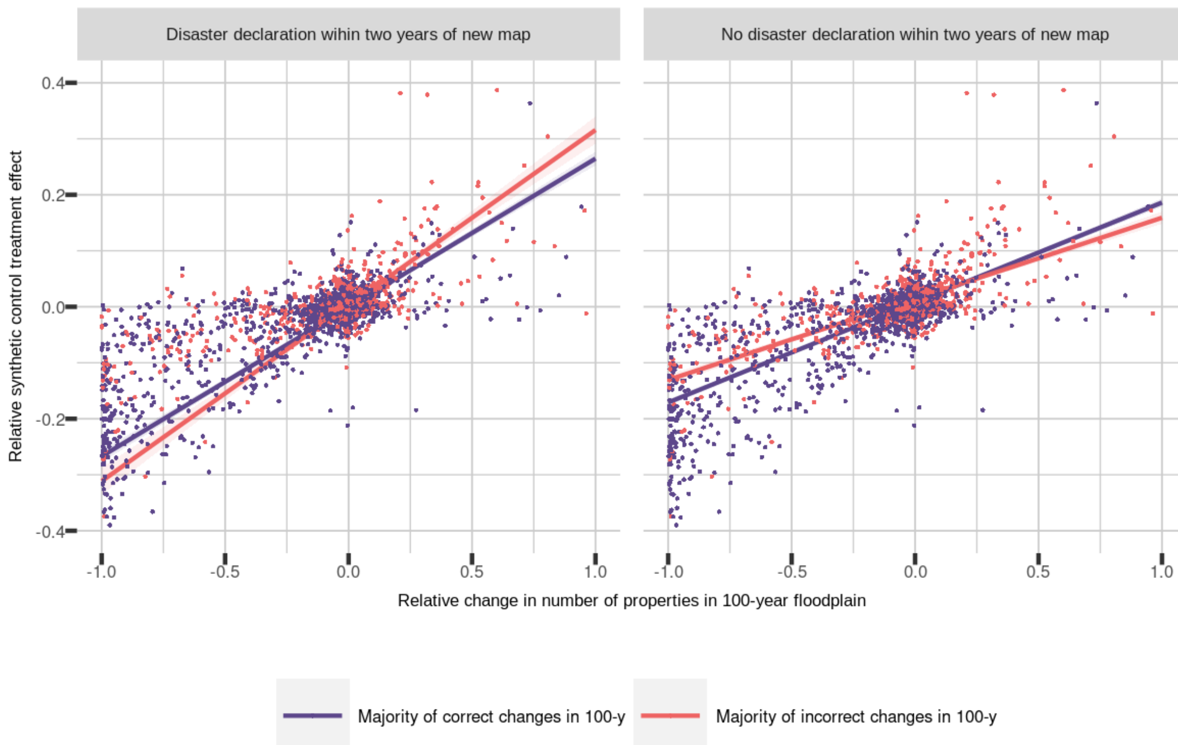

Counts are based on residential properties that are mapped in both the Q3 product and in the 2019 digital flood maps. The FSF flood model is assumed to provide the "correct" depiction of the 100-year floodplain. Additional counts for residential properties that remained correctly mapped inside or outside of the 100-year floodplains are omitted.

the number of properties in the 100-year floodplain ("treatment intensity"). The Y-axis shows the treatment effects divided by the number of total residential properties in the census tracts, and I separate treatment effects in neighborhoods where a FEMA disaster was declared within two years of the flood map. The figure further differentiates between census tracts where a majority of the rezoning was correct (purple) or incorrect (pink), and the purple and pink lines represents linear regression estimates of the treatment effects against the relative change in the number of properties rezoned inside the 100-year floodplain.

As before, we note a strong relationship between the magnitude of rezoning and the magnitude of the treatment effect: the more rezoning outside of the 100-year floodplain, the more negative the treatment effects. Furthermore, we also note that the correctness of the flood map update rezoning does not seem to be associated with a greater demand response to the map update, conditional on treatment intensity. This suggests that most of the knowledge about flood risk is conveyed by the public maps, rather than private information.

Overall, these results show that using local control units and matching on the number of pre-FIRM

Figure 1.13: Synthetic control estimates at 24 months by 100-year floodplain rezoning




Marginal effects of a change in the number of properties rezoned inside the 100-year floodplain on the synthetic control treatment effect estimates. The marginal effects are estimated separately per map correctness and disaster declaration group. Each dot represents a census tract-specific treatment effect estimate of the impact of the flood map update on the demand for flood insurance after 24 months using synthetic controls augmented by ridge regression. For each treated unit, the donor pool comprises never-treated census tracts within the same FEMA region.

properties and share of policies in the 100-year floodplain provide estimated treatment effects in striking agreements with those of section 1.4: rezoning inside the 100-year floodplain increase take-up, while rezoning outside reduce the demand for insurance. Most map-induced declines of the demand for insurance were concentrated in neighborhoods with a higher share of African Americans, a pattern primarily driven by greater rezoning of properties outside of the 100-year floodplain in these ares. Individuals appear to respond to the official flood map updates whether or not the information conveyed in these maps is correct, suggesting these maps might be the main source of information. As such, their impacts on consumer welfare could be negative.

## 1.6 A model of flood insurance demand

The impact of providing flood risk information on welfare depends not only on the quality of the new information, but also on consumer risk preferences. In this section, I develop a model of insurance demand that I calibrate with the universe of flood map updates and independent risk estimates from the First Street Foundation Flood Model to quantify the impact of map updates on welfare.

### 1.6.1 Model presentation

A homeowner $\omega$ decides whether or not to purchase insurance, with utility $u_{\omega 1}$ if she purchases and $u_{\omega 0}$ if she does not. Let $\alpha(\omega)$ denote her (absolute) risk aversion. The homeowner will purchase insurance as long as her risk risk aversion is above a threshold value[43] $k_\omega$:

$$u_{\omega 1} > u_{\omega 0} \iff \alpha(\omega) > k_\omega \tag{1.8}$$

The probability of $\omega$ purchasing insurance is then

$$\begin{aligned} p(\alpha(\omega) > k_\omega) &\iff 1 - p(\alpha(\omega) \leq k_\omega) \\ &\iff 1 - F_c(k_\omega) \end{aligned} \tag{1.9}$$

for some census tract-specific cumulative distribution function $F_c$.

The expected social surplus obtained from homeowner $\omega$ before the map update can be written as

$$\begin{aligned} W_{\omega,pre} &= p(\alpha(\omega) > k_{\omega,pre}) \cdot (CS_{w,pre} + PS_{w,pre}) \\ &= p(\alpha(\omega) > k_{\omega,pre}) \cdot (WTP_\omega - r_{\omega,pre} + r_{\omega,pre} - b_\omega) \\ &= p(\alpha(\omega) > k_{\omega,pre}) \cdot (WTP_\omega - b_\omega) \end{aligned} \tag{1.10}$$

where $CS_{\omega,pre}$ and $PS_{\omega,pre}$ are the surpluses from the insurance contract that accrue prior to the map update to the homeowner and insurer respectively,[44] $b_\omega$ is the expected (annual) cost due to flooding, $r_{\omega,pre}$ is the price of the contract prior to the map update, and $WTP_\omega$ is the homeowner's underlying willingness-to-pay for flood insurance. Intuitively, the welfare that arises from homeowner $\omega$ is 0 if she does not purchase insurance, while it is the difference between her willingness-to-pay and the expected costs of supplying the contract if she purchases insurance. Importantly, we consider her *true* willingness-to-pay, that is, the one

[43] For an individual with preferences represented by well-behaved von Neumann–Morgenstern utility functions, this threshold exists and is unique. A sufficient (but not necessary) set of conditions that guarantees existence and uniqueness of $k_\omega$ is as follows: i) $u_{\omega 1} = u_{\omega 0}$ for exactly one value of $\alpha$, ii) $lim_{\alpha \to +\infty}(u_{\omega 1} - u_{\omega 0}) \in R_{++}$, iii) $lim_{\alpha \to -\infty}(u_{\omega 1} - u_{\omega 0}) \in R_{--}$, iv) $\frac{\partial(u_{\omega 1} - u_{\omega 0})}{\partial \alpha} > 0$.

[44] These welfare definitions are net of any administrative costs of providing the contract.

that would arise *if* she had access to correct flood risk information.

Integrating over homeowners, the expected welfare at the census tract level before the map update is computed as

$$
\begin{aligned}
E[W_{c,pre}] &= \int_{\omega} (WTP_{\omega} - b_{\omega}) \cdot 1(\alpha_{\omega} > k_{\omega,pre}) \cdot dF_c \\
&= \int_{\alpha_{\omega} > k_{\omega,pre}}^{+\infty} (WTP_{\omega} - b_{\omega}) \cdot dF_c
\end{aligned}
\tag{1.11}
$$

and similarly after the map update for $E[W_{c,post}]$.

Finally, we sum over tracts to get the aggregate (consumer) welfare change:

$$
\Delta W = \sum_{c} E[W_{c,post}] - E[W_{c,pre}] \tag{1.12}
$$

Note that in this set-up, all *social* welfare changes attributable to the map update arise due to changes in the *probabilities* of households purchasing flood insurance. Map updates only impact the support of the integrals by changing the cut-off values $k_{\omega,\cdot}$, while the willingness-to-pay for insurance and the expected annual costs of flooding remain constant through the flood map update.

Assuming a constant household-specific willingness-to-pay for insurance allows us to view all improvements in the information provided to households as net welfare gains, *even if* the corrected risk information shows the household to be at increased risk of flooding. To clarify, imagine that Valentina is willing to pay \$10,000 dollars for a Marc Chagall painting, while she would be willing to pay \$0 dollars for an imitation of the same painting. If she purchases the painting for \$200 and it is later revealed that the painting is a fake, our theoretical set-up views this information revelation as (weakly) welfare increasing: Valentina experienced a net loss from the initial purchase – since all surpluses should be computed in light of her *true* willingness-to-pay – which was zero for a fake Chagall painting. Revealing correct information does not lead to further losses.

However, assuming that the expected annual costs of flooding are constant throughout the map update explicitly rules out long-term adaptation, such as elevating the property or constructing flood walls, which the household might undertake in response to new risk information. Our model therefore provides a lower-bound on the welfare gains from corrected flood maps.[45]

[45] In the context of a partial equilibrium analysis only. In a general equilibrium framework that allows for household sorting and preferences defined over endogenous neighborhood amenities, the welfare effects of correcting flood risk information are a priori ambiguous.

### 1.6.2 Model calibration

Estimating the model requires additional assumptions. For tractability, I assume that the risk preferences of individuals are captured by a Constant Absolute Risk Aversion utility function. This is consistent with evidence from the flood insurance data: wealthier neighborhoods are more likely to purchase flood insurance, which suggests that risk aversion in the context of flood risk does not decrease with income.[46] Dropping the time subscripts for clarity, equation 1.8 then becomes

$$\begin{aligned} -exp\Big(-\alpha(\omega)(Y(\omega)-r(\omega))\Big) &> -\hat{p}_\omega exp\Big(-\alpha(\omega)(Y(\omega)-L(\omega))\Big) \\ &\quad -(1-\hat{p}_\omega)exp\Big(-\alpha(\omega)(Y(\omega))\Big) \\ \iff -exp\Big(\alpha(\omega)r(\omega)\Big) &> -\hat{p}_\omega exp\Big(\alpha(\omega)L(\omega)\Big) - 1 + \hat{p}_\omega \\ \iff \alpha(\omega) &> k_\omega \end{aligned} \tag{1.13}$$

where $\hat{p}_\omega$ is the *perceived probability* of flooding, $L(\omega)$ are the *perceived* damages associated with flooding, $r(\omega)$ is the price of the insurance contract, $Y(\omega)$ is income, and $\alpha(\omega)$ is still the absolute risk aversion parameter. $L(\omega), r(\omega)$ and $Y(\omega)$ are observed by households and the econometrician, whereas $\alpha(\omega)$ is known by the household only. The cutoff value $k_\omega$ does not have a closed-form solution but can be computed numerically for each property.

The willingness-to-pay is derived as the price of the insurance contract that makes the homeowner indifferent between purchasing insurance and being uninsured:

$$\begin{aligned} -exp(\alpha(\omega)WTP_\omega) &= \int_0^{+\infty} -exp\Big(\alpha(\omega)D_d(\omega)\Big)dF_{\omega,d} \\ \iff WTP_\omega &= \frac{ln\bigg(\int_0^{+\infty} exp\Big(\alpha(\omega)D_d(\omega)\Big)dF_{\omega,d}\bigg)}{\alpha(\omega)} \end{aligned} \tag{1.14}$$

where $D_d(\omega)$ are the (true) expected damages due to flooding that occurs with inundation depth $d$ for property $\omega$, and $F_{\omega,d}$ is the probability distribution of flooding at each depth $\omega$. $D_d(\omega)$ and $F_{\omega,d}$ are both taken from the First Street Foundation Flood Model.

Consistent with the empirical evidence presented in Sections 1.4 and 1.5, I assume that households perceived probabilities come from the official FEMA flood maps: households perceive the probability of flooding to be 1% in the 100-year floodplain, 0.2% in the 500-year floodplain, and zero outside of it.[47] In

[46] This functional form assumption further allow Y to be simplified out of equation 1.13, which avoids the thorny issue of estimating income from aggregate data.

[47] This assumption will bias the impacts of new flood maps towards zero, as the analysis above demonstrates the existence of within-neighborhoods spatial spillover effects of flood maps. An alternative assumption would be to specify a parametric structure for these spillovers, for instance assuming that households impute their probability of flooding as linearly or exponentially decreasing based on their location relative to the nearest floodplain. While such alternative assumptions are plausible, the data

the normative part of the analysis below, I assess the welfare impacts of correcting the official floodplain boundaries by assuming that the 100-year and 500-year floodplains are updated to reflect the FSF Flood Model.

I further assume that households perceive the costs of flooding based on the expected damages given by the FSF Flood Model. This is a strong assumption, which will again tend to under-estimate the impacts of new flood maps. Figure D.1 in the appendix shows that results are similar if we assume instead that the perceived damages are given by the average insurance claims in the neighborhood.

To recover the price of the insurance contract for all households (including for those who do not purchase insurance), I use neighborhood-, floodplain-, and time-specific premium averages. To assess the welfare impacts of moving from the current premiums to actuarially fair premiums, I further estimate welfare changes assuming that insurance prices are provided by annual expected losses estimates in the FSF model. Additional details are provided in the appendix.

In order to integrate out the risk aversion parameters, I assume these parameters follow a Fréchet distribution within each census tract. This distribution allow for a fat upper tail and is governed by two parameters for strictly positive support:[48]

$$p(\alpha(\omega) \leq k_\omega) = exp\Big( - \gamma_c (\frac{k_\omega}{A_c})^{-\theta_c} \Big) \tag{1.15}$$

where $\theta_c$ and $A_c$ are the shape and scale parameters.

Recent work by Wagner (2021) shows that the observed low level of demand for flood insurance cannot be rationalized by risk loving homeowners, but instead reveal the existence of *frictions* that limit demand below optimal levels. While the previous sections of this paper revealed that lack of information and incorrect information about flood risks are among the decisive frictions that limit demand, my results do not rule out the existence of other frictions. This poses a challenge for the structural estimation: unless *all* existing frictions are correctly specified, one cannot separately estimate frictions and risk preferences. If one were willing to assume that incorrect information is the only friction limiting demand, then data on flood insurance transactions, together with the synthetic control estimates and a functional form assumption for the distribution of risk aversion preferences are sufficient to estimate the parameters governing the distribution of risk aversion. This approach is presented in section D.0.2 in the appendix.

Instead of assuming correct specifications of all the frictions in order to back out risk aversion parameters, my analysis proceeds by assuming the parameters governing the distribution of the risk aversion parameters and *then* estimate the impacts of providing correct information. Following the literature on insurance demand

do not allow me to estimate these parametric relationship.

[48]Consistent with the literature, I assume that all homeowners are weakly risk averse.

and Wagner (2021), in my preferred calibration I assume the distribution of risk aversion parameters has an expected value of $10^{-5}$, and I assess robustness of the findings for values of $10^{-4}$ and $10^{-6}$.[49]

Overall, 17.4 unique properties have a non zero perceived flooding probability and contribute to at least one of the welfare measures.[50] Computing expected welfare at the property level can be done sequentially and independently for all properties: first compute the homeowner-specific threshold value $k_\omega$, then integrate out the risk aversion parameters.

### 1.6.3 Structural estimates

Figure 1.14 presents several welfare metrics under a range of assumptions. Figure 1.14.A presents the empirical cumulative distribution functions of relative changes in social welfare over census tracts between the paper-based and digital maps, for three different expected value for the risk aversion parameters. A total of 20,276 census tracts experience a non-zero change in social welfare. The map updates decrease welfare in almost exactly half of these census tracts, while the other half experiences gains. These relative estimates are not sensitive to the choice of risk aversion parameters. Looking at this figure, a policy maker with distributional concerns might hope that the gains are concentrated in more disadvantaged neighborhoods, or that the welfare gains are large enough to potentially compensate the welfare losses that occur in other neighborhoods.[51] Unfortunately, the next figure reveals this is not the case.

Figure 1.14.B depicts the relative social welfare impacts of the map update, aggregated by neighborhood income decile. The pink line presents the welfare changes that occurred computed at the actual prices before and after the map update, while the black line presents the welfare changes *that would have occurred if* insurance premiums had been actuarially fair. First, note that for all income groups, the map updates *decreased* welfare: the relative changes vary between -3% and -13%, with no clear pattern with respect to the income group. Second, correcting insurance premiums to reflect actuarially fair prices would have mitigated some of these losses, but would not have led to substantial welfare gains. This reveals that insurance pricing is not the main driver behind the welfare losses – incorrect risk mapping is.

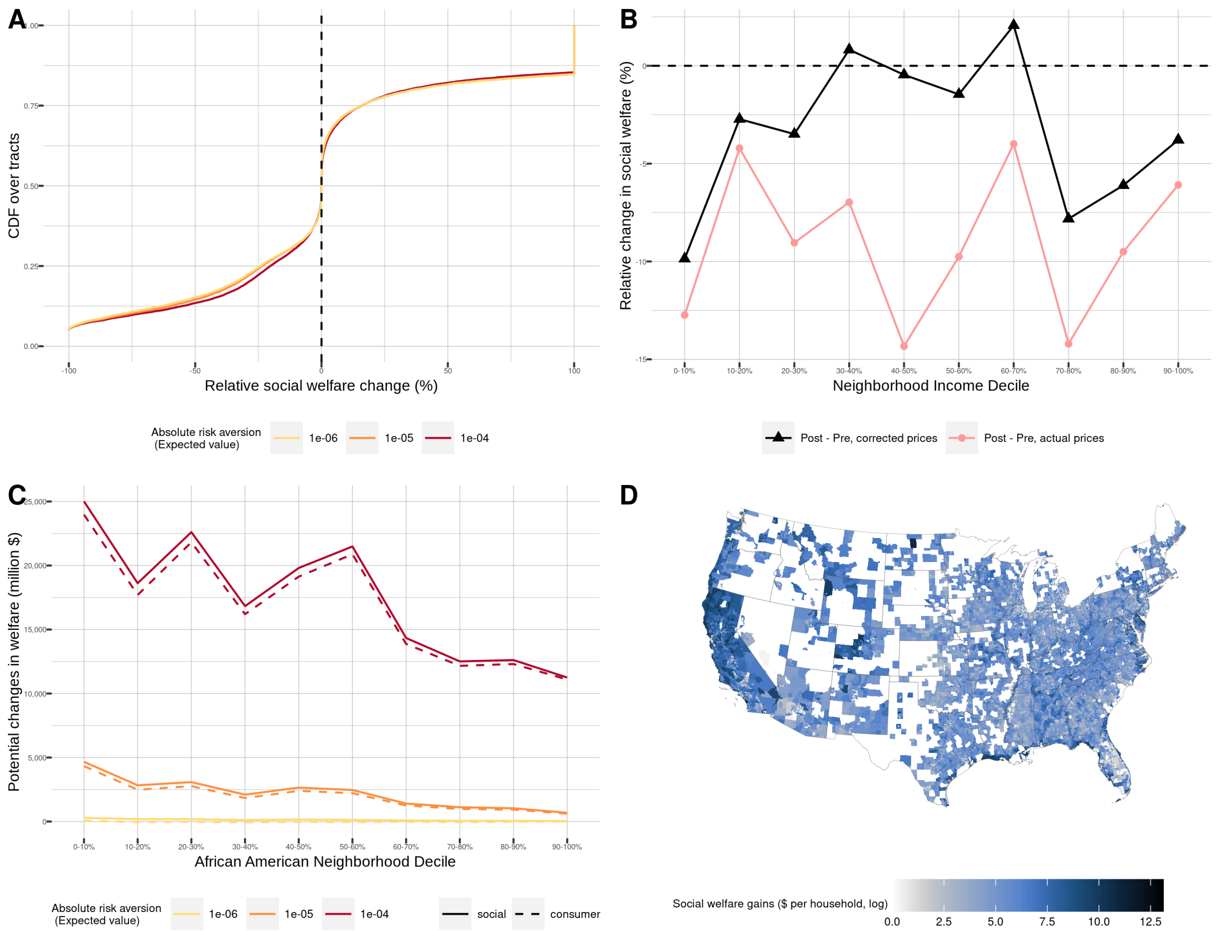
To assess the potential gains from improving flood risk mapping, Figure 1.14.C presents the welfare impacts of correcting the 100-year and 500-year floodplain boundaries pre-map update using the First Street Foundation Flood Model, aggregated by African American neighborhood decile. The solid lines depict changes in social welfare while dashed lines depict changes in consumer welfare. Amounts are reported for

[49] The shape parameter is fixed at 2 to allow for the existence of very risk aversion consumers. Other values are plausible and lead to qualitatively similar results.

[50] Welfare estimates focused on communities for which I can assign a probability of flooding pre-map update comprise 12 million unique properties.

[51] Although Sallee (2019) shows that the potential compensation criterion is intrinsically problematic, as the lack of information about *who* loses from a given policy tend to prevent the compensation of ever being implemented.

Figure 1.14: Social welfare impacts of updating flood maps




A: Cumulative distribution of social welfare changes over census tracts between the paper-based and digital map updates. Census tracts with a change of 0 are excluded, and the empirical cdf is shown on [-100%,+100%] for clarity. B: Social welfare impacts of updated flood maps aggregated by neighborhood income deciles, assuming an expected risk aversion value of $10^{-5}$. The pink line depicts relative welfare changes using true insurance premiums (pre and post map updates), while the black line assumes actuarially fair premiums before and after the map updates. C: Potential welfare gains of updating pre-modernization flood maps following the 100-year and 500-year floodplains as estimated in the FSF model, aggregated by African American neighborhood decile and assuming actuarially fair insurance pricing in corrected floodplain boundaries. Estimates for three different values of the expected risk aversion parameters are represented, along with both social and consumer welfare measures. D: Same as C, but assuming a lower risk aversion value of $10^-6$. Spatial distribution of the potential social welfare gains of updating post-modernization flood maps following the 100-year and 500-year floodplains as estimated in the FSF model, assuming actuarially fair insurance pricing in corrected floodplain boundaries. Dollar amounts are expressed per household experiencing any change in floodplain boundaries (in log and bounded at 0 for clarity).

three different expected values of the risk aversion distribution.

Figure 1.14.C shows that welfare gains would have been substantial if the map update had followed correct floodplain boundaries and updated premiums to reflect actuarially fair prices. Using a plausible parameter of $10^{-5}$ for the expected value of the risk aversion distribution, updating the maps pre-digitization to the FSF floodplain boundaries would have yielded annual gains of about 20 billion dollars, with the largest gains concentrated in white and wealthy neighborhoods. Note that while the relative changes in welfare are not

sensitive to positive values of risk aversion parameters, the absolute changes in welfare strongly depend on risk preferences. Multiplying risk aversion by a factor of ten sharply increases welfare gains, while using a risk aversion parameter of $10^{-6}$ substantially compresses any welfare gains. For all three risk aversion parameters, consumer welfare closely track social welfare. At very low expected value of risk aversion ($10^{-4}$), consumer welfare changes becomes negative, showing that more risk neutral homeowners would stand to lose the most from correcting floodplain boundaries and insurance pricing.

Finally, Figure 1.14.D shows the spatial distribution of welfare gains that would arise if we moved from the current flood maps and premiums to the corrected floodplain boundaries and actuarially fair premiums, with the welfare gains expressed per rezoned household and assuming an expected risk aversion parameter of $10^-5$. While flood-prone neighborhoods on the Atlantic coast are predicted to realize substantial welfare gains, households rezoned to a different floodplain in California, Oregon, and Washington State, as well as households residing inland near the Mississippi, Ohio and Tennessee rivers would also experience large gains in welfare.

## 1.7 Conclusion

This paper investigates the public provision of flood risk information and its impacts on the demand for flood insurance in the United States. I compile novel data on the evolution of official floodplain boundaries, the largest disaster risk mapping and information provision effort ever undertaken by a national government.

Leveraging the staggered updating of flood maps, I find that individuals' demand for insurance responds to changing risk information, but not to the mode of provision of this information: removing properties from the 100-year floodplain causes insurance demand to fall, while rezoning properties inside the 100-year floodplain increases flood insurance take-up, both inside and outside of the 100-year floodplain. These spillover effects suggest that information plays a substantial role in the demand for flood insurance. In contrast, digitizing previously-existing floodplain boundaries has insignificant impact on the demand for flood insurance, suggesting that information access costs were not limiting insurance demand.

Using a novel approach based on the local and clustered estimation of synthetic controls, I estimate that flood map updates caused 40,000 additional properties to be covered by flood insurance after two-years in areas where the 100-year floodplain was expanded, while causing more than 100,000 households to drop their insurance coverage in areas where the 100-year floodplain was shrunk. While all states experienced both negative and positive treatment effects from the modernization of flood maps, most map-induced declines in insurance coverage occurred in census tracts with a higher share of African Americans. Although flood maps rezoning reduced flood insurance premiums in these neighborhoods, declines in insurance coverage raise concerns regarding the vulnerability of communities already suffering from under-investments. The econometric approach employed in this paper identifies neighborhoods where the map updates caused the largest declines in insurance coverage, and can be used by policymakers and floodplain managers to prioritize flood risk assessments where resilience seems most jeopardized.

These findings highlight that the modernization and continued updating of flood insurance rate maps, which had a direct cost to taxpayers of more than four billion dollars in the past 15 years, directly determines the demand for flood insurance. Flood map updates between 2005 and 2020 rezoned more than 2.3 million properties outside of the 100-year floodplain, while only rezoning 1.1 million properties inside the 100-year floodplain: a net *decrease* in nominal flood risk. A comparison with a state-of-the-art model of flood risk further reveals that 5.2 million properties should have been rezoned inside the 100-year floodplain, yet were ignored. The net decrease in the number of properties zoned inside the 100-year floodplain caused an aggregate decline in both the demand for flood insurance. As such, the modernization of flood maps is failing on its promise to increase insurance coverage. Under plausible assumptions on the risk preferences of households, I find that map updates over the past fifteen years decreased social welfare.

A reconciliation bill approved by the House Financial Services Committee in September 2021 aims to provide three billion dollars to improve flood maps throughout the U.S (House Financial Services Committee, 2021). My results show that such investment has the potential to benefit consumers through improved flood risk information and increased insurance take-up while simultaneously increasing the solvency of the National Flood Insurance Program by bringing premiums closer to actuarially fair rates. However, incorrect map updates could further increase vulnerability to flood risk, as demand for insurance responds to new flood maps whether or not the information they contain is accurate. Given existing concerns regarding the substantial costs of risk mapping, the lack of capacity to keep maps up-to-date, and the potential role of homeowners lobbying, a promising avenue is to use newly-available and independent models of flood risks as a standard baseline for the new flood maps. Specific deviations from these models could be warranted and included in official products where local communities have relevant knowledge about risk or newly-developed infrastructure. But as climate change increases the frequency and magnitude of natural disasters, government-provided risk information needs to primarily rely on independent and regularly updated models that reflect the best available science.

## 1.8 References


Abadie, A. (2020). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature In Press*, 44.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science 59*(2), 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*(1), 113–132.

Abadie, A. and J. L'Hour (2019). A Penalized Synthetic Control Estimator for Disaggregated Data. *Working Paper*, 1–35.

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton (2016). The value of connections in turbulent times: Evidence from the United States. *Journal of Financial Economics 121*(2), 368–391.

Allaire, M. (2018). Socio-economic impacts of flooding: A review of the empirical literature. *Water Security 3*(May), 18–26.

Anderson, D. R. (1974). The National Flood Insurance Program. Problems and Potential. *The Journal of Risk and Insurance 41*(4), 579.

Armal, S., J. R. Porter, B. Lingle, Z. Chu, M. L. Marston, and O. E. Wing (2020). Assessing property level economic impacts of climate in the US, new insights and evidence from a comprehensive flood risk assessment tool. *Climate 8*(10), 1–20.

Atreya, A., S. Ferreira, and E. Michel-Kerjan (2015). What drives households to buy flood insurance? New evidence from Georgia. *Ecological Economics 117*, 153–161.

Barber, R. F., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2020). Predictive inference with the jackknife+. *arXiv*.

Barreca, A., K. Clay, O. Deschenes, M. Greenstone, and J. S. Shapiro (2016). Adapting to climate change: The remarkable decline in the US temperature-mortality relationship over the Twentieth Century. *Journal of Political Economy 124*(1), 105–159.

Bates, P. D., N. Quinn, C. Sampson, A. Smith, O. Wing, J. Sosa, J. Savage, G. Olcese, J. Neal, G. Schumann, L. Giustarini, G. Coxon, J. R. Porter, M. F. Amodeo, Z. Chu, S. Lewis-Gruss, N. B. Freeman, T. Houser, M. Delgado, A. Hamidi, I. Bolliger, K. E. McCusker, K. Emanuel, C. M. Ferreira, A. Khalid, I. D. Haigh, A. Couasnon, R. E. Kopp, S. Hsiang, and W. F. Krajewski (2021). Combined Modeling of US Fluvial, Pluvial, and Coastal Flood Hazard Under Current and Future Climates. *Water Resources Research 57*(2), 1–29.

Bell, H. M. and G. A. Tobin (2007). Efficient and effective? The 100-year flood in the communication and perception of flood risk. *Environmental Hazards 7*(4), 302–311.

Beltrán, A., D. Maddison, and R. J. Elliott (2018). Is Flood Risk Capitalised Into Property Values? *Ecological Economics 146*, 668–685.

Ben-Michael, E., A. Feller, and J. Rothstein (2019). Synthetic Controls and Weighted Event Studies with Staggered Adoption.

Ben-Michael, E., A. Feller, and J. Rothstein (2020). The Augmented Synthetic Control Method. WP.

Bin, O. and C. E. Landry (2013). Changes in implicit flood risk premiums: Empirical evidence from the housing market. *Journal of Environmental Economics and Management 65*(3), 361–376.

Borusyak, K. and X. Jaravel (2017). Revisiting Event Study. Working Paper.

Botzen, W. J., O. Deschenes, and M. Sanders (2019). The economic impacts of natural disasters: A review of models and empirical studies. *Review of Environmental Economics and Policy 13*(2), 167–188.

Boustan, L. P., M. E. Kahn, and P. W. Rhode (2012). Moving to higher ground: Migration response to natural disasters in the early twentieth century. *American Economic Review 102*(3), 238–244.

Bradt, J. T., C. Kousky, and O. E. Wing (2021). Voluntary purchases and adverse selection in the market for flood insurance. *Journal of Environmental Economics and Management 110*(0), 102515.

Byskov, M. F., K. Hyams, P. Satyal, I. Anguelovski, L. Benjamin, S. Blackburn, M. Borie, S. Caney, E. Chu, G. Edwards, K. Fourie, A. Fraser, C. Heyward, H. Jeans, C. McQuistan, J. Paavola, E. Page, M. Pelling, S. Priest, K. Swiderska, M. Tarazona, T. Thornton, J. Twigg, and A. Venn (2021). An agenda for ethics and justice in adaptation to climate change. *Climate and Development 13*(1), 1–9.

Cabral, M. and C. Hoxby (2015). The Hated Property Tax: Salience, Tax Rates, and Tax Revolts. *NBER Working Paper*.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics, forthcoming*, 1–31.

Chambwera, M., G. Heal, C. Dubeux, S. Hallegatte, L. Leclerc, A. Markandya, B. A. McCarl, R. Mechler, and J. E. Neumann (2014). Economics of adaptation. In Cambridge University Press (Ed.), *Climate Change 2014 Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects*, pp. 945–978. Cambridge, United Kingdom.

Chen, J. J., V. Mueller, Y. Jia, and S. K. H. Tseng (2017). Validating migration responses to flooding using satellite and vital registration data. *American Economic Review 107*(5), 441–445.

Chetty, R., A. Looney, and K. Kroft (2009). Salience and taxation: Theory and evidence. *American Economic Review 99*(4), 1145–1177.

Chivers, J. and N. E. Flores (2002). Market failure in information: The National Flood Insurance Program. *Land Economics 78*(4), 515–521.

Cutler, D. M., R. S. Huckman, and M. B. Landrum (2004). The role of information in medical markets: An analysis of publicly reported outcomes in cardiac surgery. *American Economic Review 94*(2), 342–346.

Davlasheridze, M., K. Fisher-Vanden, and H. Allen Klaiber (2017). The effects of adaptation measures on hurricane induced property losses: Which FEMA investments have the highest returns? *Journal of Environmental Economics and Management 81*, 93–114.

de Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-Way Fixed Effects Estimators With Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

Department of Homeland Security (2012). Exhibit 300 BY13 FEMA - Risk Mapping, Assessment and Planning (RISKMAP) Program.

Doudchenko, N. and G. W. Imbens (2016). Balancing, Regression, Difference-in-differences and synthetic control methods: a synthesis. *NBER Working Paper*.

Eisenack, K., S. C. Moser, E. Hoffmann, R. J. Klein, C. Oberlack, A. Pechan, M. Rotter, and C. J. Termeer (2014). Explaining and overcoming barriers to climate change adaptation. *Nature Climate Change 4*(10), 867–872.

FEMA (1996). Q3 flood data use guide. Technical Report March.

FEMA (2006). Flood Map Modernization Mid-Course Adjustment. Technical report, FEMA.

FEMA (2009). Risk Mapping, Assessment, and Planning (Risk MAP) Multi-Year Plan: Fiscal Years 2010-2014. Technical report, Homeland Security.

FEMA (2016). Flood Insurance Study - City of New Orleans and Orleans Parish, Louisiana. Technical report, FEMA.

First Street Foundation (2020). First Street Foundation Flood Model (FSF-FM) Technical Documentation. Technical report, First Street Foundation.

Gall, M., B. Boruff, and S. L. Cutter (2007). Assessing Flood Hazard Zones in the Absence of Digital Floodplain Maps: Comparison of Alternative Approaches. *Natural Hazards Review 8*(1), 1–12.

Gallagher, J. (2014). Learning About an Infrequent Event: Evidence from Flood Insurance Take-Up in the US. *American Economic Journal: Economic Policy 6*(3), 206–233.

Gibson, M. and J. T. Mullins (2020). Climate Risk and Beliefs in New York Floodplains. *Journal of the Association of Environmental and Resource Economists 7*(6), 1069–1111.

Gobillon, L. and T. Magnac (2016). Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls. *Review of Economics and Statistics 98*(3), 535–551.

Goodman-Bacon, A. (2021). Difference-In-Differences with Variation in Treatment Timing. *Journal of Econometrics 225*(2), 254–277.

Graff Zivin, J., Y. Liao, and Y. Panassié (2019). How Hurricanes Sweep Up Housing Markets: Evidence from Florida. *NBER Working Paper*.

Hallegatte, S. (2014). Economic Resilience: definition and measurement. Technical Report May, The World Bank.

Hallegatte, S., M. Fay, and E. B. Barbier (2018). Poverty and climate change: introduction. *Environment and Development Economics 23*(03), 217–233.

Highfield, W. E., S. A. Norman, and S. D. Brody (2013). Examining the 100-Year Floodplain as a Metric of Risk, Loss, and Household Adjustment. *Risk Analysis 33*(2), 186–191.

Hinkel, J., D. Lincke, A. T. Vafeidis, M. Perrette, R. J. Nicholls, R. S. Tol, B. Marzeion, X. Fettweis, C. Ionescu, and A. Levermann (2014). Coastal flood damage and adaptation costs under 21st century sea-level rise. *Proceedings of the National Academy of Sciences of the United States of America 111*(9), 3292–3297.

Hino, M. and M. Burke (2021). The effect of information about climate risk on property values. *Proceedings of the National Academy of Sciences 118*(17).

Horn, D. P. and B. W. Webel (2019). Introduction to the national flood insurance program (NFIP). Technical report.

Hornbeck, R. and S. Naidu (2014). When the levee breaks: Black migration and economic development in the American South. *American Economic Review 104*(3), 963–990.

House Financial Services Committee (2021). Ammendment in the nature of a substitute to the committee print offered by Ms. Waters of California, Title IV Subtitle A — Creating and Preserving affordable, equitable and accessible housing for the 21st Century.

Hu, Z. (2022). Social interactions and households' flood insurance decisions. *Journal of Financial Economics 144*(2), 414–432.

Hudson, P., W. J. Botzen, L. Feyen, and J. C. Aerts (2016). Incentivising flood risk adaptation through risk based insurance premiums: Trade-offs between affordability and risk reduction. *Ecological Economics 125*, 1–13.

Kahn, M. E. (2021). Adapting to Climate Change: Markets and the Management of an Uncertain Future.

Kennedy, P. W., B. Laplante, and J. Maxwell (1994). Pollution policy: The role for publicly provided information.

Knowles, S. G. and H. C. Kunreuther (2014). Troubled waters: The national flood insurance program in historical perspective. *Journal of Policy History 26*(3), 327–353.

Koerth, M. (2017). It's Time To Ditch The Concept Of '100-Year Floods'. *FiveThirtyEight*.

Konar, S. and M. A. Cohen (1997). Information as regulation: The effect of community right to know laws on toxic emissions. *Journal of Environmental Economics and Management 32*(1), 109–124.

Kousky, C. (2014). Informing climate adaptation: A review of the economic costs of natural disasters. *Energy Economics 46*, 576–592.

Kousky, C. (2016). Flood Insurance: Why Don't People Buy It? Technical report.

Kousky, C., H. Kunreuther, B. Lingle, and L. Shabman (2018). The Emerging Private Residential Flood Insurance Market in the United States. (July), 1–51.

Kousky, C., B. Lingle, H. Kunreuther, and L. Shabman (2019). Moving the Needle on Closing the Flood Insurance Gap. Technical Report 2013, Wharton University of Pennsylvania.

Kousky, C., B. Lingle, and L. Shabman (2017). The Pricing of Flood Insurance. *Journal of Extreme Events 04*(02), 1750001.

Kunreuther, H. C., M. V. Pauly, and S. McMorrow (2013). Insurance and behavioral economics: Improving decisions in the most misunderstood industry.

Landry, C. and D. Turner (2020). Risk perceptions and flood insurance: insights from homeowners on the georgia coast. *Sustainability 12*(24), 1–19.

Lea, D. and S. Pralle (2022). To appeal and amend: Changes to recently updated Flood Insurance Rate Maps. *Risk, Hazards and Crisis in Public Policy 13*(1), 28–47.

Liao, Y. and P. Mulder (2021). What's at Stake? Understanding the Role of Home Equity in Flood Insurance Demand. *Working Paper*, 1–51.

Ludy, J. and G. M. Kondolf (2012). Flood risk perception in lands "protected" by 100-year levees. *Natural Hazards 61*(2), 829–842.

Marsooli, R., N. Lin, K. Emanuel, and K. Feng (2019). Climate change exacerbates hurricane flood hazards along US Atlantic and Gulf Coasts in spatially varying patterns. *Nature Communications 10*(1), 1–9.

Mennis, J. (2017). Dasymetric Mapping. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, 1–10.

Michel-Kerjan, E., S. Lemoyne de Forges, and H. Kunreuther (2012). Policy Tenure Under the U.S. National Flood Insurance Program (NFIP). *Risk Analysis 32*(4), 644–658.

Michel-Kerjan, E. O. (2010). Catastrophe Economics: The National Flood Insurance Program. *Journal of Economic Perspectives 24*(4), 165–186.

Morrissey, W. A. (2008). FEMA Funding for Flood Map Modernization. Technical report, Congressional Research Service.

Moser, S. C. and J. A. Ekstrom (2010). A framework to diagnose barriers to climate change adaptation. *Proceedings of the National Academy of Sciences of the United States of America 107*(51), 22026–22031.

Mulder, P. (2019). Dynamic Adverse Selection in Flood Insurance. WP. pp. 1–45.

Mulder, P. (2021). Mismeasuring Risk: The Welfare Effects of Flood Risk Information. *Working paper*.

National Reearch Council (2015). Affordability of National Flood Insurance Program Premiums. Technical report.

Nguyen, C. N. and I. Noy (2020). Measuring the impact of insurance on urban earthquake recovery using nightlights. *Journal of Economic Geography 20*(3), 857–877.

NOAA (2020). Billion dollar weather and climate disasters.

Office of Inspector General (2017). FEMA Needs to Improve Management of Its Flood Mapping Programs. Technical report.

Ouazad, A. and M. E. Kahn (2021). Mortgage Finance and Climate Change : Securitization Dynamics in the Aftermath of Natural Disasters. *NBER Working paper*.

Petrolia, D. R., C. E. Landry, and K. H. Coble (2013). Risk preferences, risk perceptions, and flood insurance. *Land Economics 89*(2), 227–245.

Pope, J. C. (2008). Do seller disclosures affect property values? Buyer information and the hedonic model. *Land Economics 84*(4), 551–572.

Pralle, S. (2019). Drawing lines: FEMA and the politics of mapping flood zones. *Climatic Change 152*(2), 227–237.

Robbins, M. W., J. Saunders, and B. Kilmer (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association 112*(517), 109–126.

Sallee, J. M. (2014). Rational inattention and energy efficiency. *Journal of Law and Economics 57*(3), 781–820.

Sallee, J. M. (2019). Pigou Creates Losers: on the Implausibility of Achieving Pareto Improvements from Efficiency Enhancing Policies. WP. *NBER Working Paper Series*.

Shao, W., S. Xian, N. Lin, H. Kunreuther, N. Jackson, and K. Goidel (2017). Understanding the effects of past flood events and perceived and estimated flood risks on individuals' voluntary flood insurance purchase behavior. *Water Research 108*, 391–400.

Shi, L., E. Chu, I. Anguelovski, A. Aylett, J. Debats, K. Goh, T. Schenk, K. C. Seto, D. Dodman, D. Roberts, J. T. Roberts, and S. D. Van Deveer (2016). Roadmap towards justice in urban climate adaptation research. *Nature Climate Change 6*(2), 131–137.

Shr, Y.-H. and K. Y. Zipp (2019). The Aftermath of Flood Zone Remapping: The Asymmetric Impact of Flood Maps on Housing Prices. *Land Economics 95*(2), 174–192.

Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy 69*(3), 213–225.

Sun, L. and S. Abraham (2021). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *Journal of Econometrics 225*(2), 175–199.

The United Nations Environment Programme (2021). Adaptation Gap Report 2020. Technical report, The United Nations.

The World Bank and the Government of Mexico (2012). Improving the Assessment of Disaster Risks to Strengthen Financial Resilience. Technical report.

Tobin, R. J. and C. Calfee (2005). The National Flood Insurance Program's Mandatory Purchase Requirement: Policies, Processes, and Stakeholders. Technical Report March, American Institutes for Research, Washington, D.C.

Troy, A. and J. Romm (2004). Assessing the price effects of flood hazard disclosure under the California Natural Hazard Disclosure Law (AB 1195). *Journal of Environmental Planning and Management 47*(1), 137–162.

Turnham, J., K. Burnett, C. Martin, T. McCall, R. Juras, and J. Spader (2011). Housing Recovery on the Gulf Coast - Phase II (Results of Property Owner Survey in Louisiana, Mississippi, and Texas). Technical report.

Union of Concerned Scientists (2017). When Rising Seas Hit Home. pp. 51.

Wagner, K. (2021). Adaptation and Adverse Selection in Markets for Natural Disaster Insurance. *American Economic Journal: Economic Policy, forthcoming*.

Watkins, R. R. (2012). Gastrointestinal infections in the setting of natural disasters. *Current Infectious Disease Reports 14*(1), 47–52.

Weill, J. A., M. Stigler, O. Deschenes, and M. R. Springborn (2021). Researchers' Degrees-of-Flexibility and the Credibility of Difference-in-Differences Estimates: Evidence From the Pandemic Policy Evaluations. *NBER Working Paper*, 1–72.

Westra, S., H. J. Fowler, J. P. Evans, L. V. Alexander, P. Berg, F. Johnson, J. L. Kendon, G. Lenderink, and N. M. Roberts (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics 52*, 522–555.

Wilson, M. T. and C. Kousky (2019). The Long Road to Adoption: How Long Does it Take to Adopt Updated County-Level Flood Insurance Rate Maps? *Risk, Hazards and Crisis in Public Policy 10*(4), 403–421.

Wing, O., P. D. Bates, A. M. Smith, C. C. Sampson, K. A. Johnson, J. Fargione, and P. Morefield (2018). Estimates of present and future flood risk in the conterminous United States. *Environmental Research Letters 13*(3), 034023.

Wing, O. E., W. Lehman, P. D. Bates, C. C. Sampson, N. Quinn, A. M. Smith, J. C. Neal, J. R. Porter, and C. Kousky (2022). Inequitable patterns of US flood risk in the Anthropocene. *Nature Climate Change 12*(2), 156–162.

Wittenberg, A. (2017). The myth of the 100-year flood. *EE News, GreenWire*.

Wright, J. K. (1936). A method of mapping densities of population. *Geogrpahical Review 26*(1), 103–110.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

# Appendices

# Appendix A

# Data preparation

### A.0.1 Timing and content of the flood map updates

The flood insurance rate maps are provided at the community level. The community is usually a county, but can also be a city, a town, a borough or a parish. Previous work focused on flood maps relied on the "Community Status Book" maintained by FEMA,[1], which records the effective dates of the communities' first FIRM (paper-based flood map) and current FIRM or DFIRM (where the D stands for *Digital*, i.e., post map update). However, because the Community Status Book only records the date of the first FIRM and current (D)FIRM, it cannot be used to identify transitions from a FIRM to a DFIRM, and from a DFIRM to a more recent DFIRM, unless the community only ever received one DFIRM – and these communities are *not* identified in the Community Status Book. In addition, although most census tracts can be easily assigned to an NFIP community (in particular when the community is a county which received a unique county-wide map), several census tracts are instead covered by a community-specific DFIRM and cannot be directly assigned a map update dates.

To circumvent these issues, I use new data on digital yearly snapshots of the National Flood Hazard Layer since 2012 to assign DFIRM effective dates at the census tract-level. Finding, assembling, and cleaning these files was a substantial undertaking: most files were not maintained by FEMA and were retrieved from the GIS archives of contractors who worked with FEMA to produce these maps. I intersect the census tracts polygons with the DFIRM effective dates polygons and record the effective date indicated at the census tract centroid. I then compare the intersected DFIRM identification code with the census tract FIPS code: if the DFIRM was issued for an entire county, then the DFIRM ID and the first five-digits of the census tract FIPS code must match. For communities that are not entire counties (for instance independent Virginian

[1] https://www.fema.gov/flood-insurance/work-with-nfip/community-status-book

cities), I only match on the State FIPS code and confirm the effective dates obtained with the Community Status Book. This procedure induces some measurement error on treatment timing for census tracts that are not entirely included in an NFIP community, since the effective date at the centroid of the census tract might differ from the date at which other parts of the census tract received their (D)FIRM. However, this measurement error only affects a minority of census tracts in sparsely populated communities, which themselves account for a minority of NFIP insurance purchases. Omitting these census tracts from the analysis do not noticeably impact the model estimates. Finally, for census tracts covered by a digital flood map in the 2012 National Flood Hazard Layer, I manually check that this was indeed the community's first digital flood map by comparing the digitzation date with FEMA's Map Service Center repository[2]. This platform maintains a collection of pdf files depicting historical mapping products.

To investigate the evolution of floodplain boundaries throughout the map update process, I use the Q3 data product, which depicts the floodplain boundaries of the paper-based in 2005. As described in the main text, the Q3 maps were designed to aid floodplain managers and city officials with disaster response activities, but were not widely available to the public. They however provided a digital version of floodplain boundaries as shown on the historical (paper-based) flood maps.

### A.0.2 Note on intersecting floodplains and census tracts polygons

Intersecting the different flood zones within each census tract involves the pair-wise intersection of hundreds of thousands of polygons. This process is computationally expensive, but performance can be dramatically improved by first cropping the flood zone polygons at the county level and then run on parallel cores the intersections between the cropped polygon and the census tracts located within this county.

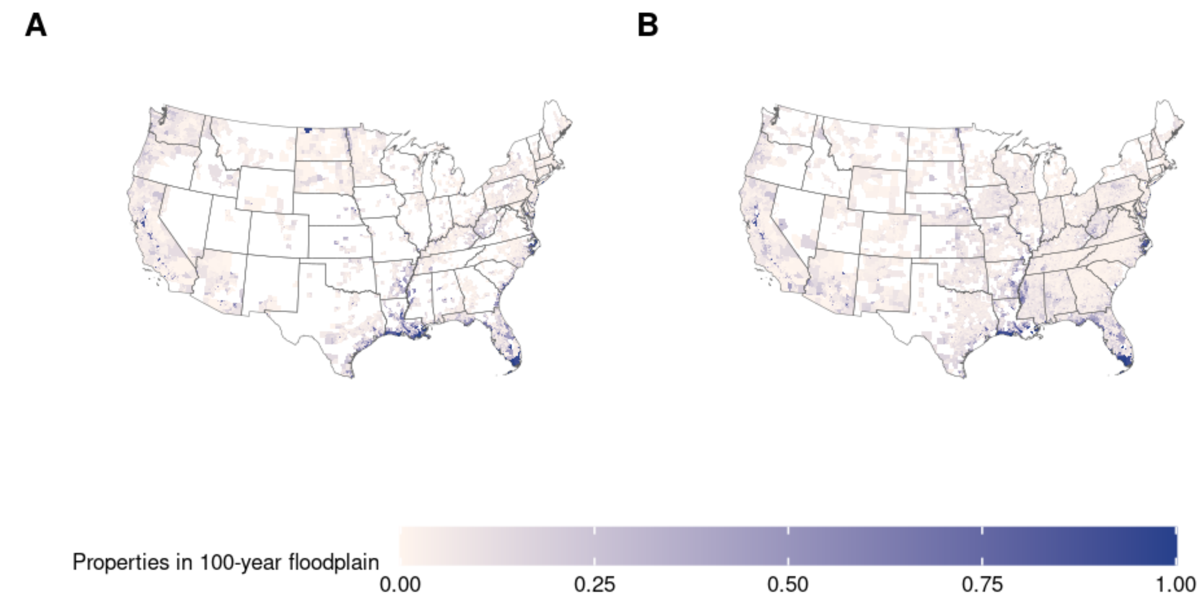
Unless a sufficient buffer is used, flood zones that barely "touch" the boundary of a census tract will induce additional intersections that are computationally costly. On the flip side, using a large buffer will induce measurement error in the computed areas. Using a minimal buffer and post-processing the intersected polygons resulted in the best performance overall. On a virtual machine with 48 vCPUS and 512 GB of RAM, all intersections can be ran within two days.

### A.0.3 Computing tract-level population and buildings count changes within different floodzones

Each census tract covers between 1,200 and 8,000 inhabitants. In very-densely populated areas, the spatial extent of a census tract is typically small, and the (spatial) share of the census tract in different flood zones

[2] https://msc.fema.gov/portal/advanceSearch

Figure A.1: Share of properties located in the 100-year floodplain




**A**: Census tract share of residential properties located in the 100-year floodplain based on the Q3 product. **B**: Census tract share of residential properties located in the 100-year floodplain 2019.

is a good approximation of the share of the census tract *population* that lives within each flood zone. This is not necessarily true in medium-sized or large census tracts, where population can often be located completely outside of any floodplains. To properly estimate population exposure to flood risk, it is therefore crucial to account for the spatial distribution of population *within* census tracts. I do so in three different ways.

First, I use the entire repository of 125 millions geolocalized residential properties maintained by the First Street Foundations and downloaded from their API in April 2021.[3] This measure of population distribution presents three advantages: (i) it captures residential properties only, which makes it a suitable measure of flood risk exposure to study how flood zone changes impact the demand for *residential* flood insurance, (ii) it is consistent with the other variables that I use from First Street to estimate flood risk exposure and expected annual average losses, and (iii) its coverage is more comprehensive than the often-used ZTRAX data (for comparison, I was only able to geolocalize 95 million residential properties from ZTRAX — I discuss the ZTRAX data below). I then intersect these residential properties with the Q3 flood maps, the 2012 DFIRMs, and the 2019 DFIRMs. I then aggregate the relevant changes in properties counts at the census tract-level, based on the effective date of the first DFIRM: for instance for a tract in Broward County (FL) that received its first (and unique) digital flood map in 2014, I compute the tract-level number of properties in each flood zone in the Q3 data as well as in the NFHL 2019 data, and take their difference to get the *change* in the number of properties located within this flood zone. I consider three different flood

[3]Due to the rate limitations of the API, the full download takes about 3 weeks of un-interrupted running time.

zones: the 100-year floodplain (or SFHA), which comprises all A and V flood zones; the 500-year flood zone (part of the X zones), and the levee-protected flood zone.

Because the First Street Foundation residential properties data is obtained from local tax assessors offices, coverage in some specific areas can be limited. When the number of residential FSF-properties in a given census tract is too low to be representative,[4] I use the 2010 dasymetric allocation of population provided by EPA's EviroAtlas. This product uses land use cover and terrain slopes to reallocate the 2010 census tract population within census tracts at the 30m pixel level.[5] By using the 2010 dasymetric raster layer, I can also ensure that any changes in the number of people living in census tract/floodplain is driven by changes in the definition of the floodplain, and *not* by population sorting and new construction following the flood map updates. The dasymetric layer covers the entire contiguous United States, but does not differentiate well between residential properties and businesses. To transform the dasymetric-population count in a flood zone into a property count, I divide the population count by 2.5 (the average number of people living in an American household).

Figure A.2 plots the changes in the number of FSF residential properties zoned in the 100-year floodplain (relative to the census tract number of properties) against the changes in the population in the 100-year zone using the dasymetric allocation. For most census tracts the two measures are in great agreement.

Finally, one might worry that these different measures of changes in flood risk exposure could conflate changes in flood zone definition with new constructions occurring disproportionately more in parcels that were rezoned outside of the 100-year floodplain. As mentioned above, the comparison with the dasymetric-population layer mitigates these concerns, but several thousands of census tracts received a flood map between 2007 and 2010. To ensure that my census tract-level measures of flood zone exposure are not driven by new properties being built in rezoned areas, I use Zillow's ZTRAX Assessment data. These data contain a large number of variables on individual structures, including their construction date. I restrict the data to geolocalized residential properties that were built *prior* to 2007. I then proceed similarly than with the FSF and dasymetric data by intersecting the properties with the Q3, the NFHL 2012 and the NFHL 2019 flood zones. Figure A.3 shows that this measure of changes in the 100-year floodplain is extremely similar to the FSF and dasymetric-derived measures presented previously.

[4] I do not use the FSF data to compute flood risk exposure changes in census tracts where the number of FSF-properties is less than 40 and where the American Community Survey estimates of the population is greater than 4 times the number of properties. In practice, this concerns less than 1,500 census tracts out of 73,745.

[5] Dasymetric mapping has a long history in geographical sciences (Wright, 1936) but remains mostly underappreciated by economists. See (Mennis, 2017) for a review of this technique.

Figure A.2: Comparison of the relative change in the number of properties in the 100-year floodplain using the FSF residential properties and the dasymetric population layer

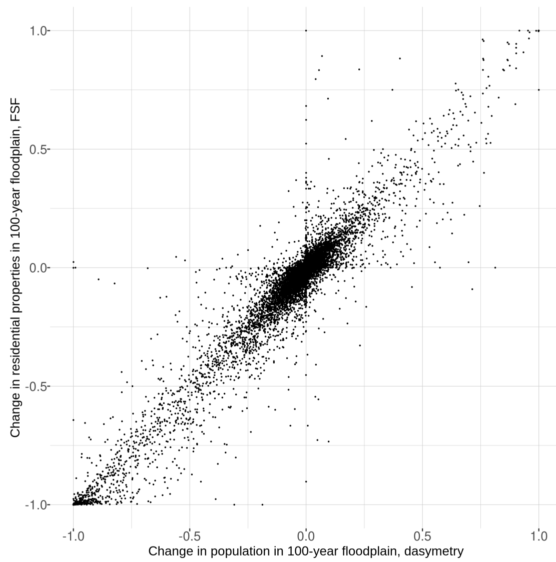



Figure A.3: Comparison of the relative change in the number of residential properties built prior to 2007 in the 100-year floodplain (using the ZTRAX data) and the dasymetric population layer

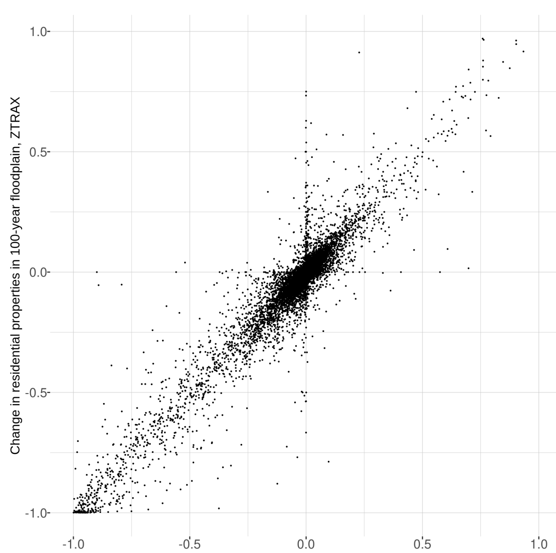
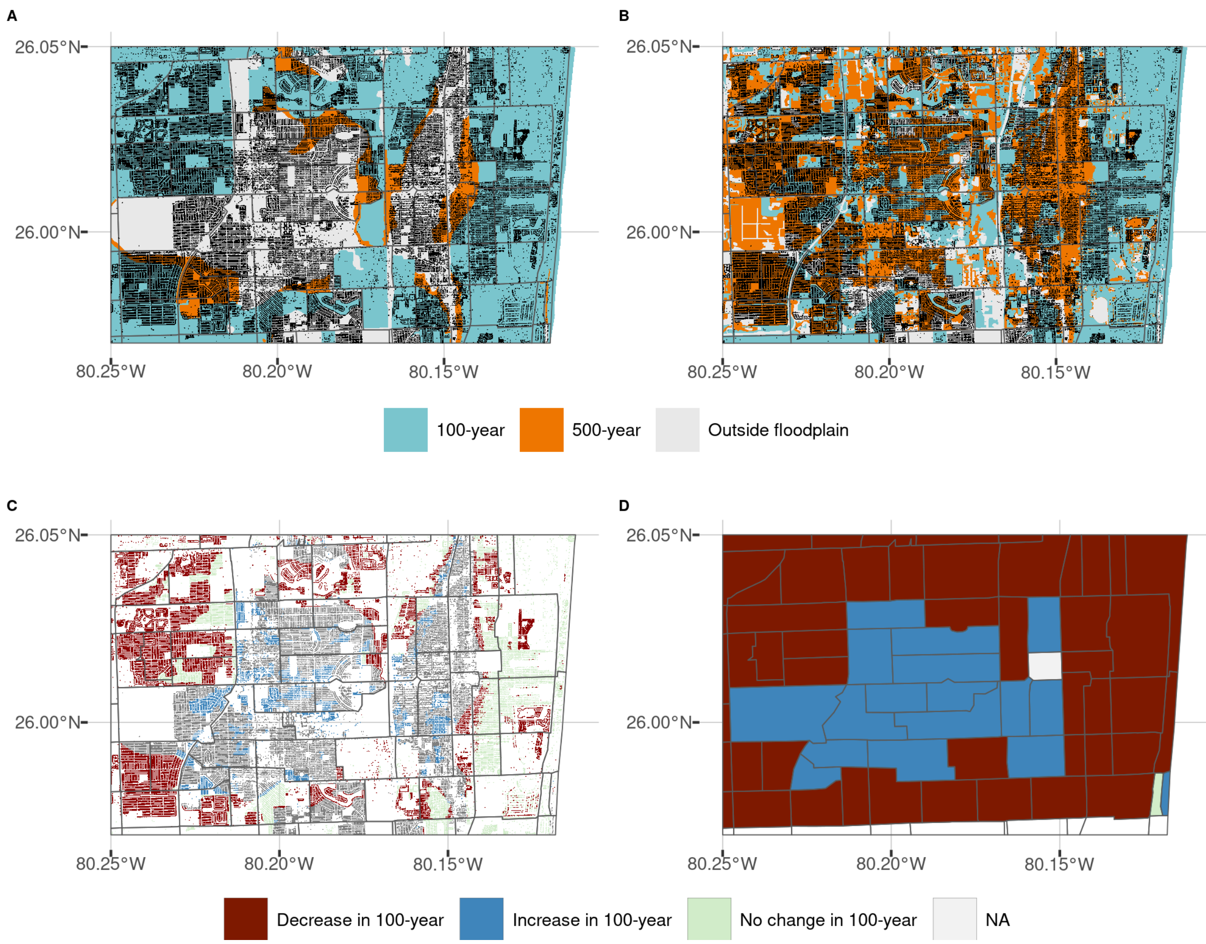


Figure A.4: Successive definition of the floodplains in South-Eastern Broward County (FL).




**A**: Flood map that was effective in 1992 (using the Q3 data).**B**: New flood map after modernization in August 2014. Black dots show the location of residential properties. **C**: Changes in the 100-year floodplain classification computed at the property-level. **D**: Aggregated changes and classification of the change at the census tract-level.

Figure A.5: Summary of the different datasets used

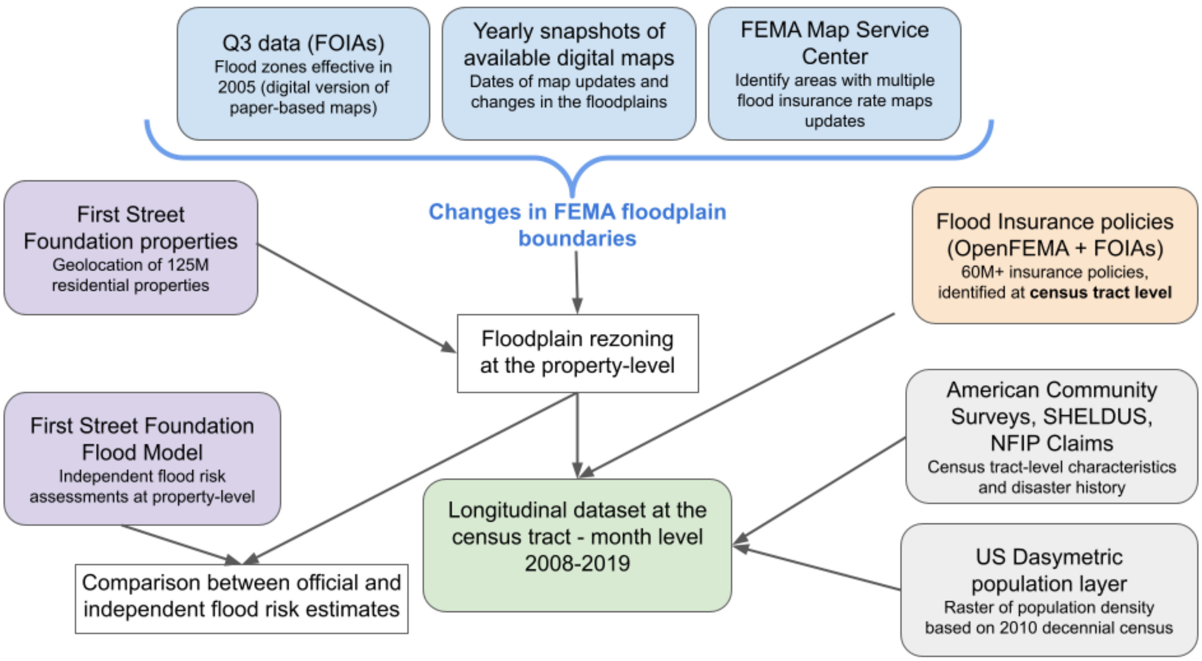
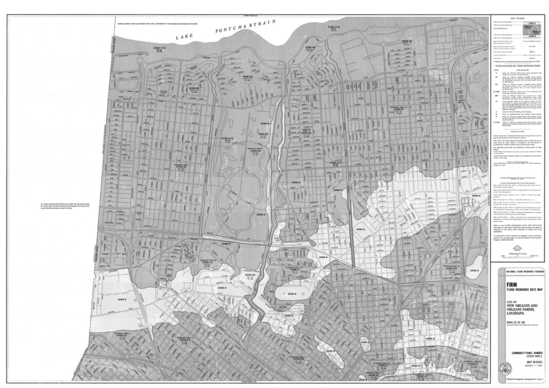



Figure A.6: Section of New Orleans' (LA) FIRM effective in 1984



### A.0.4 Panel of insurance policies at the census tract-level

I used records on insurance policies publicly provided on the OpenFEMA platform for the period 2010-2019, and records obtained through Freedom of Information Act requests for the period 1983-2010. After multiple quality checks, it was found that records prior to 2008 were missing a wide number of insurance policies – given the implausibility of the missing-at-random assumption in this context, my analysis focuses on 2008-2019 period. I exclude a small number of census tracts where the number of insurance policies was strikingly different in December 2009 and January 2010, as this demonstrates inaccuracies in the comparability of records between the two time periods. Following additional robustness checks, I found that at least part of the insurance records in California and Minnesota were incorrectly duplicated, and further excluded these records from the analysis.

Because the smallest geographical unit reported in the NFIP data is the census tract, I use this level as the panel cross-sectional unit. The choice of temporal unit, however, implies an important trade-off between granularity and effective variation. Policies can be started and terminated at any time during the year, but the distribution of the starting dates is far from being uniform within a year: in fact, policies tend to start during the summer months. As a result, keeping track of only the *year* in which a policy was started would induce substantial measurement error. On the other hand, the majority of policies are active for exactly 12 months, and most owners decide whether or not to renew their policy only once every 365 days. As a result, creating a panel at the daily level would generate a large amount of auto-correlation between observations. Finally, flood map updates occur on specific months; I therefore choose this level as the temporal unit of the panel.

The distribution of the starting dates of insurance policies is not perfectly uniform within a month: more insurance policies tend to start on exactly the 1st, the 15th/16th, or on the very last day of the month. In order to account for these variations, I specify the first month of activity of a policy either as the month in which it becomes effective *if* the policy became effective before the 15th, or the next month if the policy became effective after the 16th. For policies that become effective exactly on the 15th or 16th of the month, I assign their first month of activity based on the results of Bernouilli draws. I use a similar procedure for the last month of activity. In practice, these data cleaning choices do not impact the sign or magnitude of any of my estimates. However, they do help to reduce variations that arise purely out of measurement error.

# Appendix B

# Additional descriptive statistics and robustness checks

Table B.1: Census tracts summary statistics by year of the digital flood insurance rate map.

| Treatment year | N | Policies 2010 | SFHA share 2010 | Policies 2010 per property | Med. income | Relative change in 100-year | Area | Density | Share Black | Has Q3 data |
|---|---|---|---|---|---|---|---|---|---|---|
| 2005 | 5635 | 58 (184) | 0.34 (0.34) | 0.04 (0.11) | 33937 (12379) | -0.01 (0.05) | 43 (210) | 1220 (1211) | 0.14 (0.21) | 0.91 |
| 2006 | 3234 | 48 (228) | 0.36 (0.34) | 0.02 (0.08) | 34901 (14600) | -0.01 (0.05) | 41 (160) | 1257 (2293) | 0.15 (0.2) | 0.77 |
| 2007 | 5992 | 60 (174) | 0.24 (0.31) | 0.04 (0.11) | 33411 (15077) | 0 (0.11) | 41 (215) | 5013 (8407) | 0.22 (0.27) | 0.46 |
| 2008 | 5576 | 48 (176) | 0.35 (0.34) | 0.03 (0.12) | 32529 (12484) | -0.01 (0.09) | 59 (253) | 1578 (2527) | 0.16 (0.24) | 0.76 |
| 2009 | 7114 | 62 (197) | 0.36 (0.35) | 0.04 (0.13) | 32481 (12254) | -0.01 (0.12) | 69 (339) | 1362 (2072) | 0.15 (0.21) | 0.77 |
| 2010 | 6047 | 23 (90) | 0.35 (0.34) | 0.02 (0.09) | 32348 (13871) | 0 (0.04) | 129 (596) | 1082 (1842) | 0.13 (0.21) | 0.66 |
| 2011 | 4107 | 24 (83) | 0.42 (0.36) | 0.03 (0.12) | 29037 (10420) | -0.01 (0.07) | 121 (352) | 705 (1038) | 0.12 (0.19) | 0.56 |
| 2012 | 3836 | 43 (140) | 0.41 (0.35) | 0.03 (0.11) | 30462 (11057) | 0 (0.1) | 97 (309) | 1062 (1523) | 0.16 (0.26) | 0.66 |
| 2013 | 1148 | 38 (101) | 0.36 (0.33) | 0.02 (0.08) | 32685 (13248) | -0.01 (0.04) | 96 (298) | 785 (1405) | 0.12 (0.19) | 0.61 |
| 2014 | 2079 | 212 (401) | 0.47 (0.38) | 0.12 (0.21) | 31604 (12085) | -0.15 (0.32) | 55 (189) | 1201 (1302) | 0.15 (0.21) | 0.82 |
| 2015 | 1342 | 76 (181) | 0.47 (0.34) | 0.06 (0.14) | 30136 (10217) | -0.1 (0.25) | 100 (495) | 837 (1084) | 0.17 (0.26) | 0.79 |
| 2016 | 1074 | 71 (188) | 0.37 (0.36) | 0.05 (0.13) | 33288 (11651) | -0.02 (0.08) | 76 (455) | 1142 (1145) | 0.24 (0.29) | 0.89 |
| 2017 | 952 | 236 (384) | 0.39 (0.36) | 0.17 (0.26) | 31820 (10286) | -0.09 (0.24) | 66 (227) | 1158 (1040) | 0.15 (0.21) | 0.87 |
| 2018 | 384 | 95 (227) | 0.35 (0.34) | 0.05 (0.12) | 28726 (9554) | -0.01 (0.06) | 94 (242) | 816 (889) | 0.06 (0.13) | 0.76 |
| 2019 | 718 | 68 (295) | 0.36 (0.35) | 0.09 (0.24) | 35601 (11578) | -0.01 (0.07) | 57 (328) | 1157 (971) | 0.08 (0.14) | 0.9 |
| | 4217 | 81 (260) | 0.43 (0.36) | 0.06 (0.18) | 29762 (10713) | | 561 (3505) | 570 (933) | 0.07 (0.15) | 0.64 |

Census tracts with a flood insurance rate map prior to 2005 re-coded as 2005 for conciseness

Table B.2: Property counts in census tracts that received two map updates

| Flood zone | Q3 | NFHL 2012 | NFHL 2019 | First Street Model |
|---|---|---|---|---|
| Inside 100-year floodplain | 1,200,000 | 1,100,000 | 900,000 | 1,600,000 |
| Outside 100-year floodplain | 16,000,000 | 16,100,000 | 16,300,000 | 15,600,000 |

The data are restricted to census tracts entirely mapped in each data product.

### B.0.1 Selection into treatment and event studies

Table B.3: Selection into treatment timing and implementation

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Treatment year | | Share rezoned inside 100-year f.p. | |
| | (1) | (2) | (3) | (4) |
| Population density (IHS) | $-0.148^{**}$ (0.075) | $-0.236^{***}$ (0.062) | $-0.011^{**}$ (0.004) | $-0.004^{***}$ (0.001) |
| Median income (IHS) | $-0.869^{***}$ (0.325) | $-0.930^{***}$ (0.320) | $-0.004$ (0.008) | 0.0002 (0.004) |
| Share African Americans (IHS) | 0.266 (0.882) | 0.491 (0.850) | $-0.016$ (0.030) | $-0.050^{**}$ (0.025) |
| Disaster declaration prior to treatment | | | 0.034 (0.026) | 0.020 (0.016) |
| Average Annual Loss (IHS) | 0.093 (0.060) | | $-0.001$ (0.001) | |
| Insurance policies/property, 2008 (IHS) | | $4.727^{***}$ (1.511) | | $-0.487^{***}$ (0.147) |
| Treatment year | | | $-0.004^{**}$ (0.002) | $-0.002^{**}$ (0.001) |
| Fixed Effects | FEMA Region | FEMA Region | FEMA Region | FEMA Region |
| Sample | Positive | Positive | Positive | Positive |
| Mean outcome | 2009.3 | 2009.3 | -0.023 | -0.023 |
| Observations | 36,851 | 37,006 | 26,343 | 26,343 |
| $R^2$ | 0.080 | 0.099 | 0.068 | 0.233 |

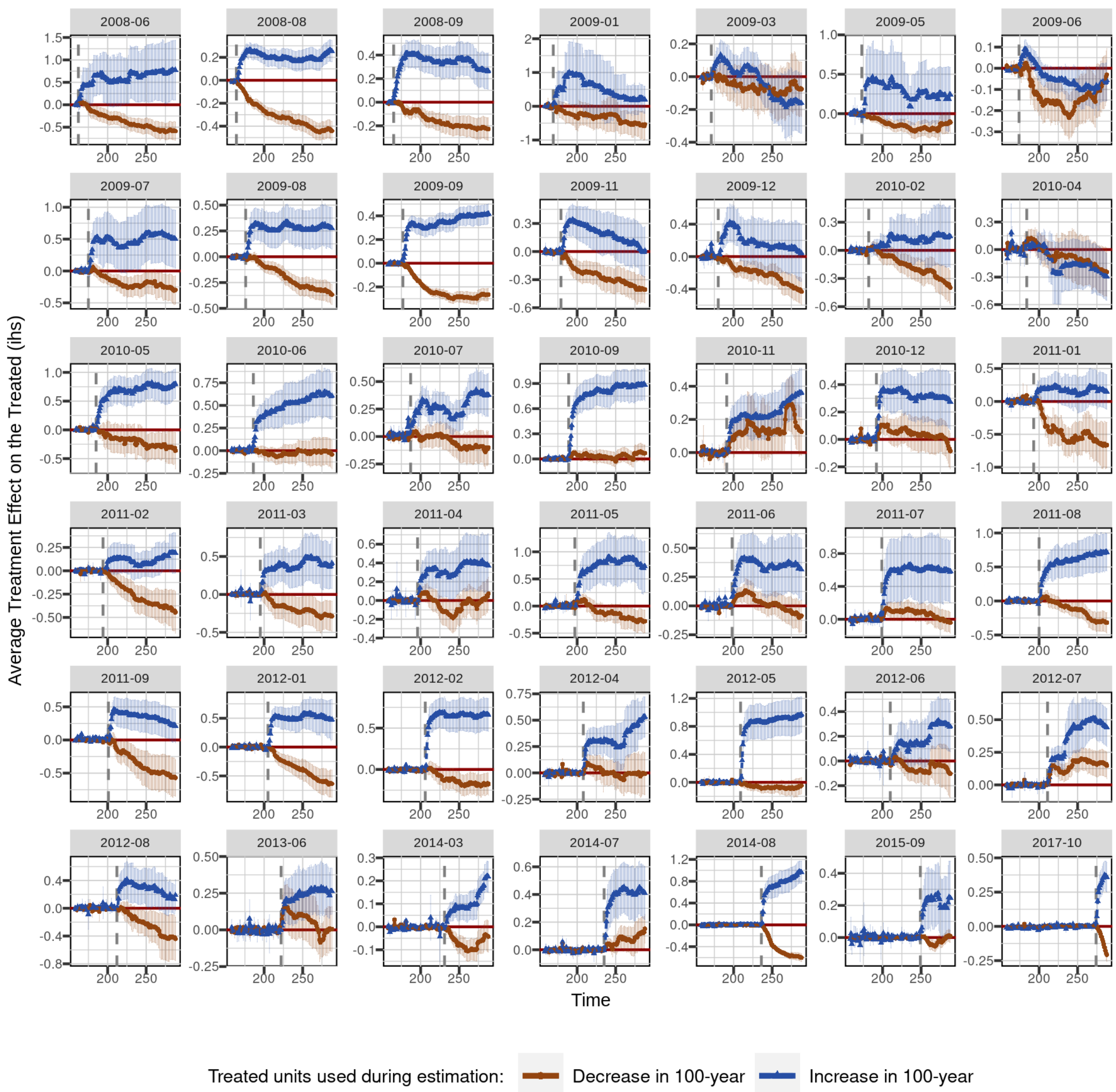
*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

"IHS" independent variables are transformed using the inverse hyperbolic sine

The "Positive" sample includes tracts ever observed with at least one insurance policy

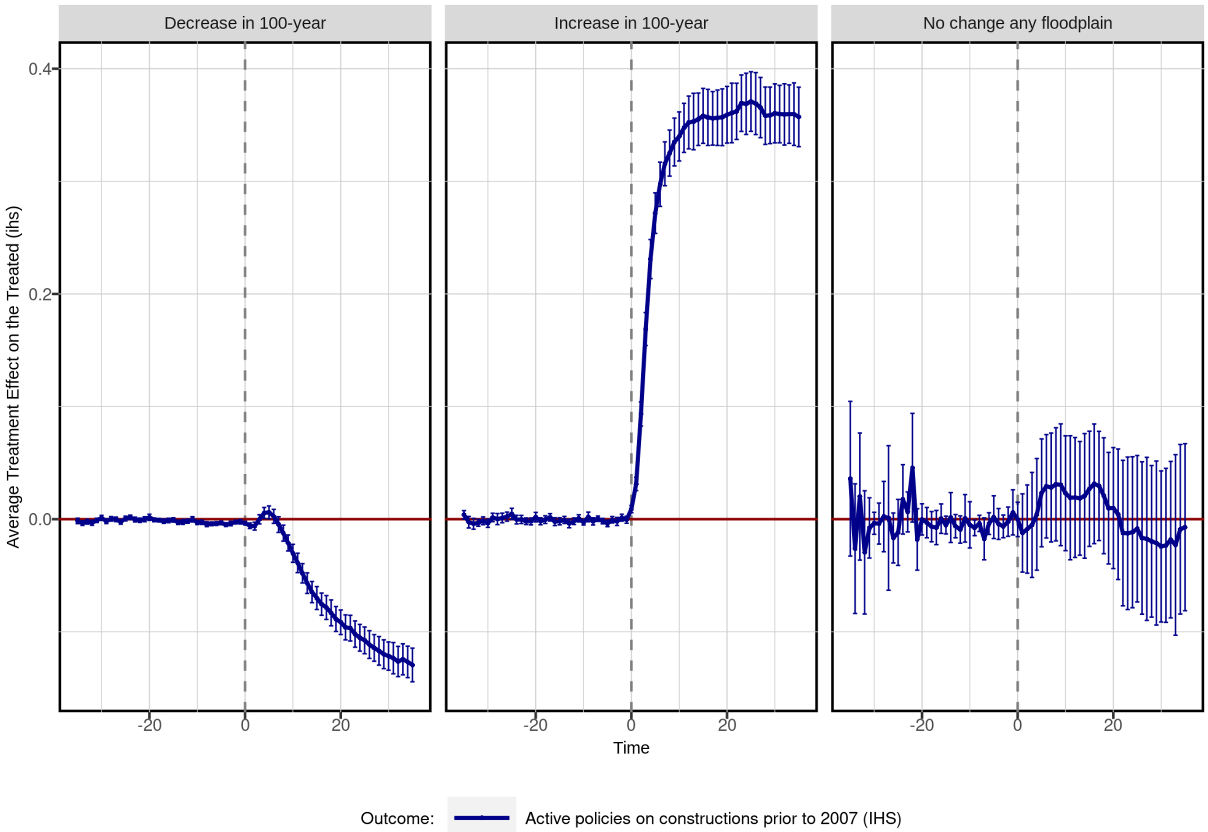
Standard errors clustered at the county level

Figure B.1: Cohort-specific event study estimates of the impacts of flood map updates on the demand for flood insurance



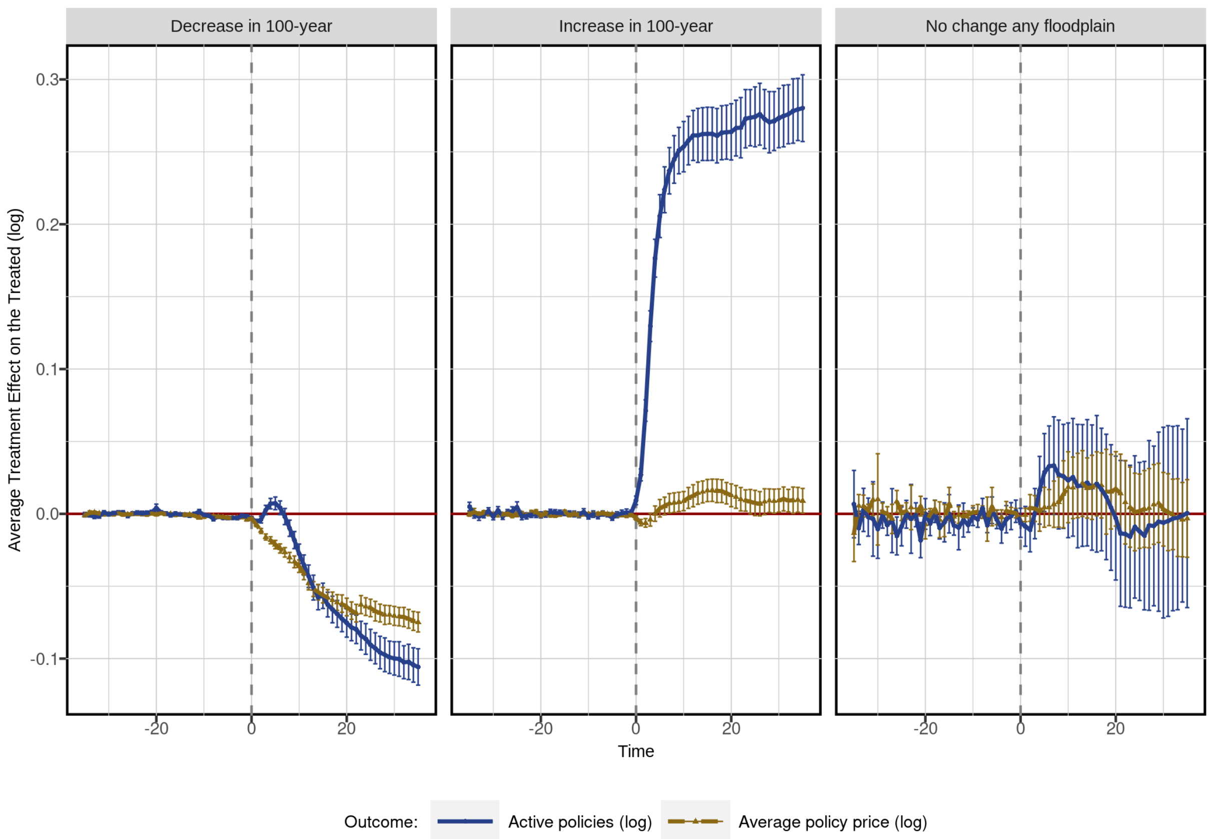

Each facet presents estimates of the average treatment effect on the treated of the impact of flood map update on insurance demand for a specific cohort (defined by the year and month of treatment). Within each cohort, the event-study are estimated separately using census tracts where the updated flood map increased (blue) or decreased (brown) the number of residential properties in the 100-year floodplain by more than 1% relative to the total number of residential properties in the census tract. The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. Error bars represent 95% confidence intervals using the multiplier bootstrap. For clarity, only the 42 largest cohorts are represented (out of 114 cohorts).

Figure B.2: Aggregated event study estimates of the impacts of flood map updates on the demand for flood insurance, robustness tests using early constructions properties



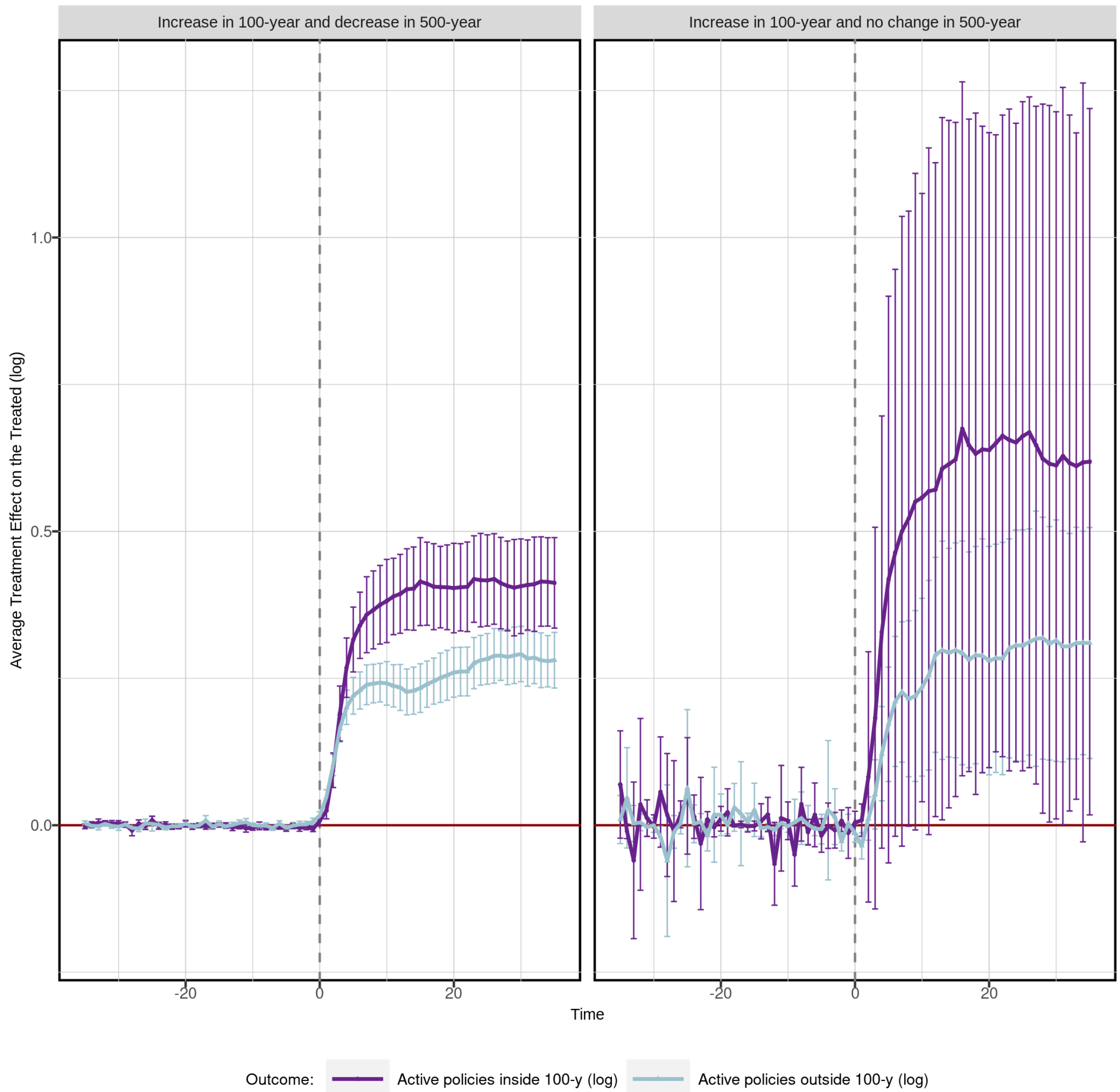

The outcomes displayed are the number of active insurance policies covering properties that were constructed prior to 2008 (solid blue line), where the construction date is taken from the insurance data. Each facet represents average treatment effects for a different treated group, using treated census tracts where the flood map update increased, decreased, or did not change the number of properties zoned inside the 100-year and 500-year floodplains (first, second and third facet respectively). The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. The outcome variables are transformed using the Inverse Hyperbolic Sine (IHS). Error bars represent 95% confidence intervals using the multiplier bootstrap.

Figure B.3: Aggregated event study estimates of the impacts of flood map updates on the demand for flood insurance and average policy price, robustness tests on sample used, functional form and covariates




Each facet represents average treatment effects for a different treated group, using treated census tracts where the flood map update increased or decreased the number of residential properties in the 100-year floodplain by more than 1% relative to the total number of residential properties in the census tract (first and second panel respectively). The third panel presents estimates for the treated census tracts that contained residential properties in the 100-year and 500-year floodplains, but where the map update did not change the number of properties located inside these floodplains. The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. The outcome variables are log-transformed. Error bars represent 95% confidence intervals using the multiplier bootstrap. The estimation sample focuses on census tracts with at least 5 insurance policies active at all times. All models control linearly for the occurrence of FEMA disaster declarations within the previous year.

Figure B.4: Aggregated event study estimates of the impacts of flood map updates on the demand for flood insurance at the census tract and flood zone-level, robustness tests on sample, transformation and covariates



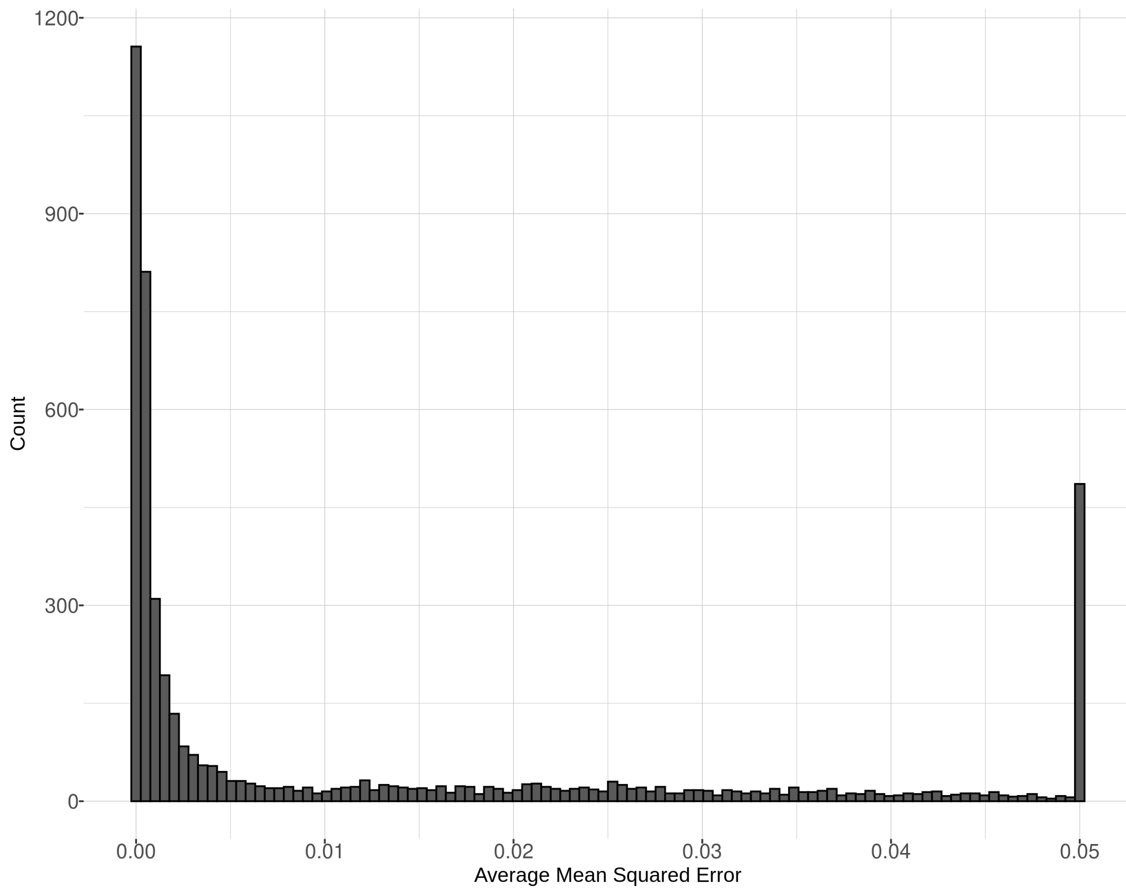

Each facet represents average treatment effects for a different treated group, using treated census tracts where the flood map update increased the number of properties in the 100-year floodplain while simultaneously rezoning properties outside of the 500-year floodplain (left) or not rezoning any properties in the 500-year floodplain (right). The outcomes are measured at the *census tract / flood zone-level*, where the flood zone designations are derived from the insurance policy data. As such, they provide the number and cost of policies that are active within each zone at the time when the policy is purchased. The control groups are comprised of census tracts that have not yet received a digital flood map at the time of treatment. The outcome variables are log-transformed. Error bars represent 95% confidence intervals using the multiplier bootstrap. The estimation sample focuses on census tracts with at least 5 insurance policies active at all times. All models control linearly for the occurrence of FEMA disaster declarations within the previous year.

# Appendix C

# Synthetic controls

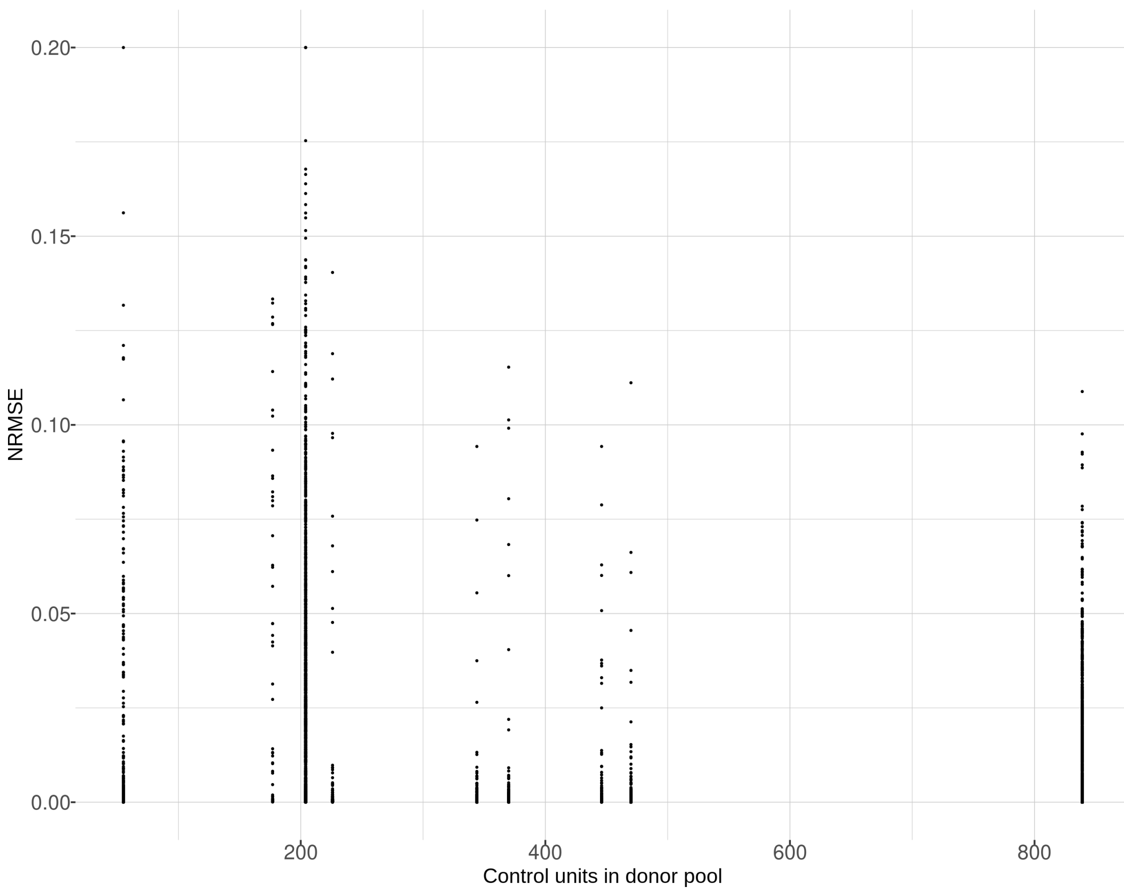
### C.0.1 Additional synthetic control results

Figure C.1: Normalized root-mean squared errors, synthetic controls



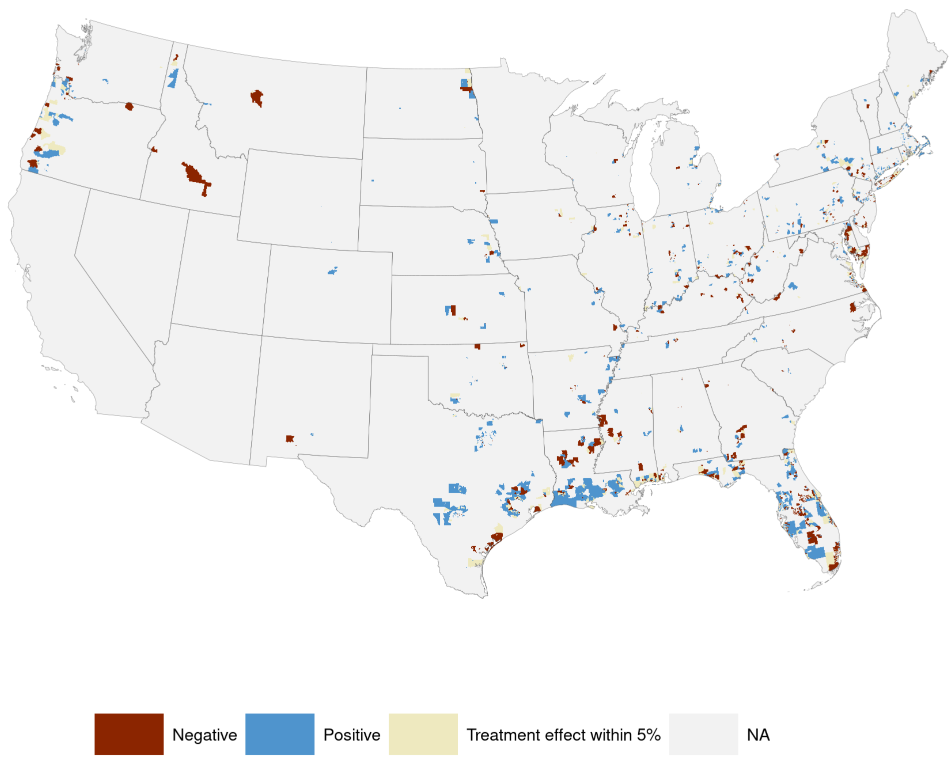

Histogram of normalized root mean squared errors between each individual tract and its (augmented) synthetic control. Synthetic controls were estimated for tracts with at least 20 insurance policies at all time, observed at least 24 months post-treatment. Matching was performed on pre-treatment outcome, share of post-FIRM policies in the tract, and share of policies in the 100-year floodplain.

Figure C.2: Normalized root-mean squared errors and number of units in synthetic control donor pool



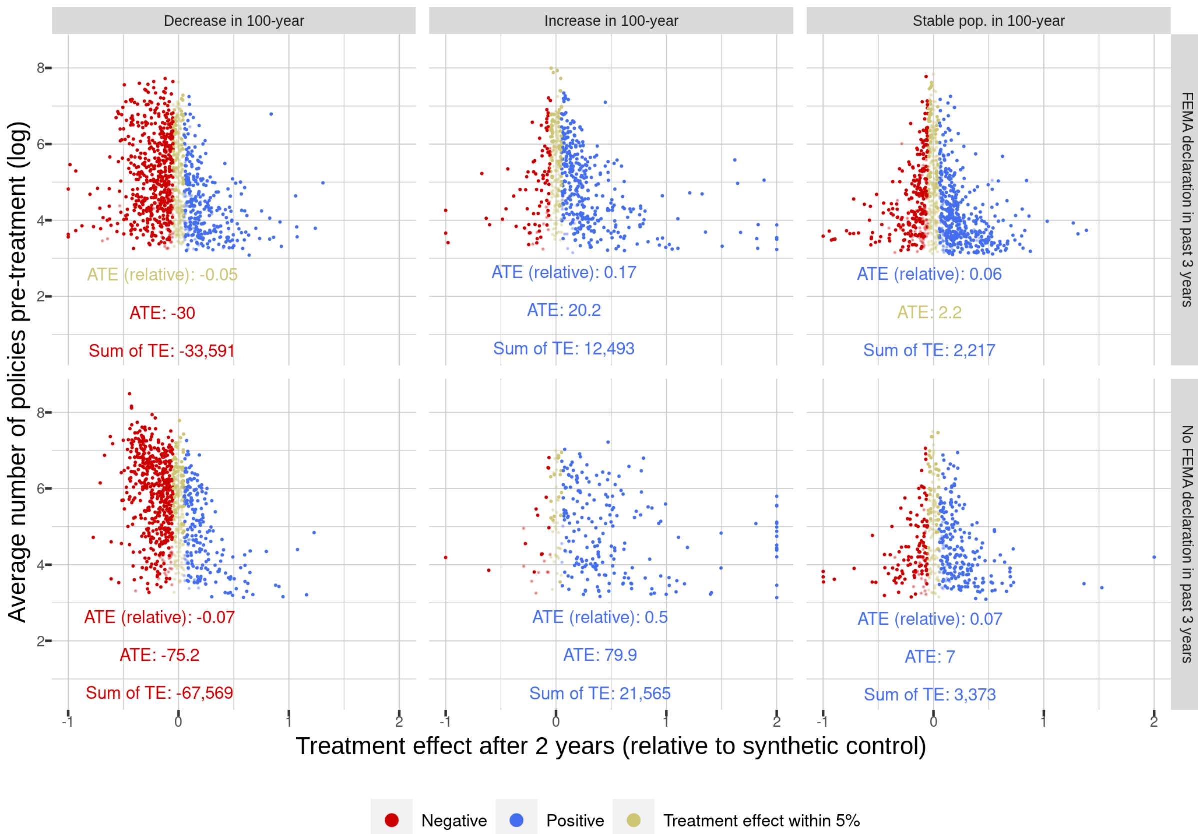

Synthetic controls were estimated for tracts with at least 20 insurance policies at all time, observed at least 24 months post-treatment. Matching was performed on pre-treatment outcome, share of post-FIRM policies in the tract, and share of policies in the 100-year floodplain.

Figure C.3: Treatment effect of the flood map update on the demand for flood insurance after two years, synthetic controls




Synthetic controls were estimated for tracts with at least 20 insurance policies at all time, observed at least 24 months post-treatment. Matching was performed on pre-treatment outcome, share of post-FIRM policies in the tract, and share of policies in the 100-year floodplain.

Figure C.4: Treatment effect of the flood map update on the demand for flood insurance after two years, by disaster declaration history, synthetic controls




Synthetic controls were estimated for tracts with at least 20 insurance policies at all time, observed at least 24 months post-treatment. Matching was performed on pre-treatment outcome, share of post-FIRM policies in the tract, and share of policies in the 100-year floodplain.

# Appendix D

# Welfare estimates

### D.0.1 Structural estimation using perceived damages from local claims

### D.0.2 Backing out risk aversion parameters from the data

The approach I employ in the paper assumes that the parameters governing the distribution of risk aversion preferences are *known*. For instance, my preferred specification fixes the shape and scale parameters of the Fréchet distribution such as to obtain an expected risk aversion value of $10^{-5}$, consistent with the literature (with a shape parameter fixed at 2 to allow for a fat upper tail, although other values are plausible). Here I show that conditional on (i) knowing the functional form of this distribution and (ii) assuming that information frictions are the only distortions constraining demand, it is possible to recover the parameters governing the distribution of risk preferences. While assumption (ii) of "no omitted frictions" seems implausible in the setting of federally-provided flood insurance, it might be credible in other settings.

Aggregated to the tract level, the share of homeowners who purchase a contract before the map update is

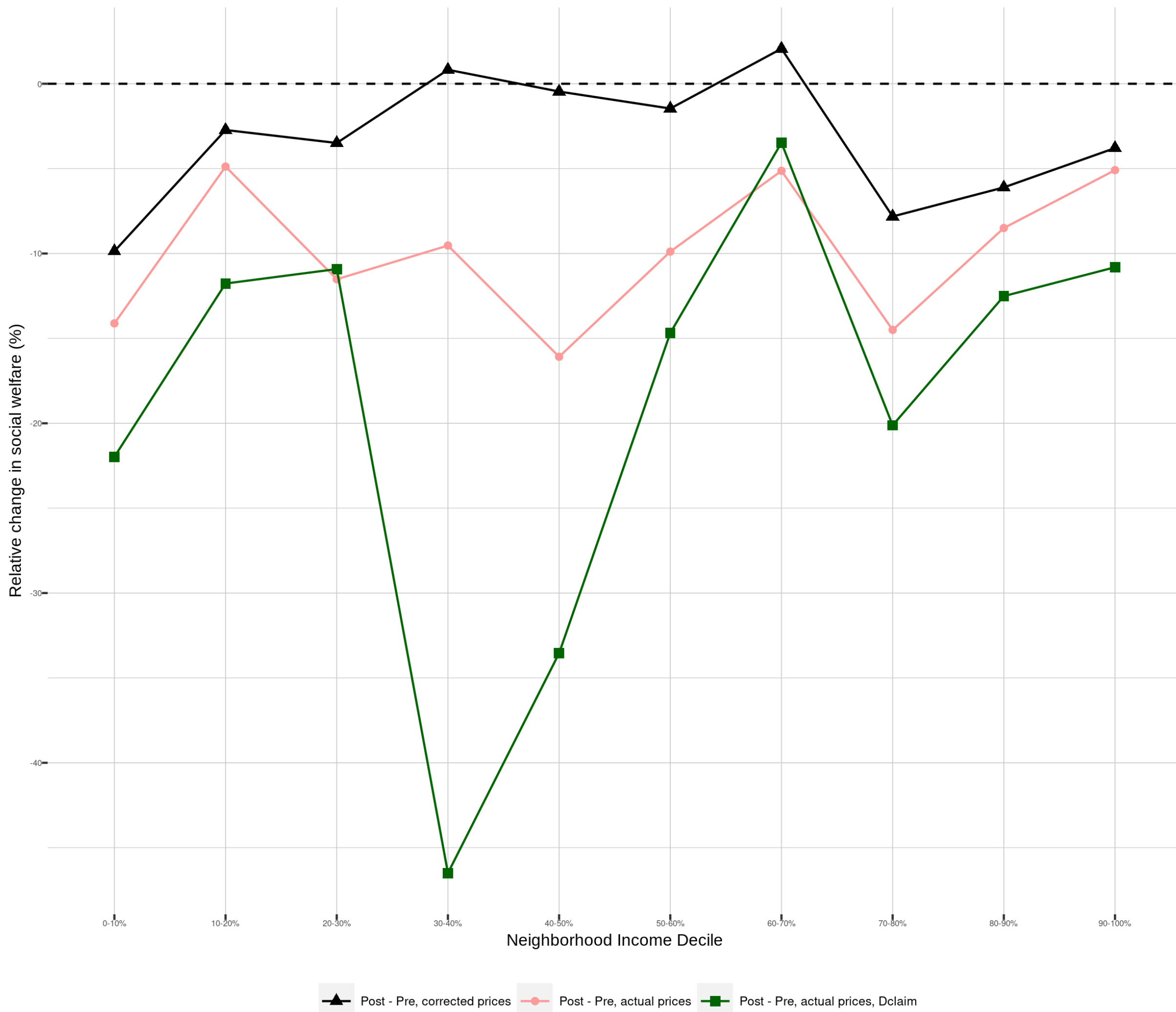
$$s_{c,pre} = \frac{\sum_{\omega} p(\alpha(\omega) > k_{\omega,pre})}{N} \tag{D.1}$$

and the share of individuals who purchase the contract after the map update is

$$\begin{aligned} \hat{s}_{c,post} &= \frac{\sum_{\omega} p(\alpha(\omega) > k_{\omega,post})}{N} \\ \hat{s}_{c,post} &= s_{c,pre} + \hat{\tau}_c \end{aligned} \tag{D.2}$$

where N is the number of residential properties in the census tract, and $\hat{\tau}_c$ is the tract-specific treatment effect of the updated flood map on insurance demand (previously estimated with synthetic controls). $s_{c,pre}$

Figure D.1: Consumer welfare impacts of updating flood maps, robustness




Consumer welfare impacts of updated flood maps aggregated by neighborhood income deciles, assuming an expected risk aversion value of $10^{-5}$. The pink line with circles depicts relative welfare changes using true insurance premiums (pre and post map updates), the black line with triangles assumes actuarially fair premiums before and after the map updates, while the green line with squares assumes that households perceive flood damages from historical claims.

is directly observed from the data, whereas $\hat{s}_{c,post}$ is the *predicted* share of individuals purchasing insurance after the map update, and where all changes in demand relative to $s_{c,pre}$ are caused by the map update.

Together with equations D.1 and D.2, this gives rise to a collection of systems of two equations (one system per census tract) which can be numerically solved to obtain $\{A_c^*, \theta_c^*\}$, the tract-specific mean risk aversion and dispersion parameters.

# Chapter 2

# Income and Mobility Inequalities at the Onset of the COVID-19 Pandemic

**Published as Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. Joakim A. Weill, Matthieu Stigler, Olivier Deschenes, and Michael R. Springborn. Proceedings of the National Academy of Sciences, 2020**

## 2.1 Introduction

In response to the threat of COVID-19, national and local governments around the world have declared emergencies, promoted safer-at-home orders, and required business closures to increase social distancing and reduce the risk of transmission. While social distancing during the 2009 H1N1 swine flu pandemic was effective in reducing infections, indirect evidence from a single region suggested this response was most pronounced in higher socioeconomic level households Springborn et al. (2015). In this paper, we use anonymized location pings data from mobile devices covering the entire U.S. to provide direct evidence of systematic differences in social distancing behavior across income levels during the COVID-19 pandemic. We show that social distancing following state emergency declarations is substantial overall, but dramatically increases in intensity with income. This finding is consistent across social distancing metrics from three difference sources of mobile device data.

There is an urgent need to identify if and to what extent lower income communities are systematically exposed to greater COVID-19 risk, especially if such effects are to be ameliorated before the full toll has accrued. Rapidly growing unemployment in a system of predominantly employer-based benefits —most critically health insurance—adds to the urgency Bitler et al. (2020). More generally, lower-income communities already experience worse health-outcomes Aizer and Currie (2014) and have a lower capacity to cope with economic and health shocks. The combination of fragile access to healthcare, especially through less secure ties with the labor market, imply a double burden of COVID-19: lower income communities appear to be most vulnerable to the economic and health impacts of the disease (e.g., due to less access to healthcare Berchick et al. (2019) and pre-existing health conditions Centers for Disease Control and Prevention (2020)) *and* here we show that they also exhibit less of the social distancing that could buffer against it.

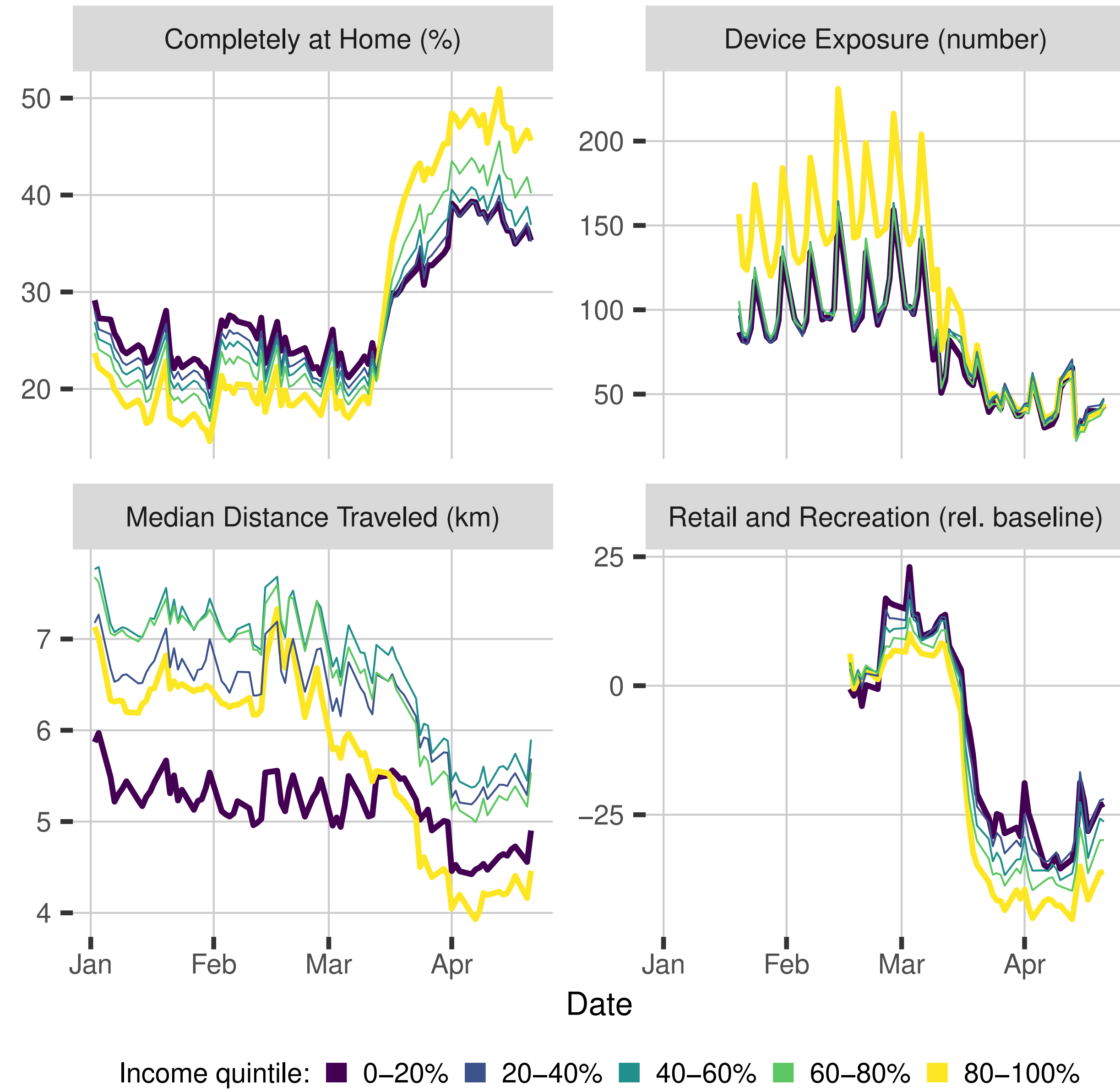
Unpacking the mechanisms through which income is associated with behavioral responses to social distancing policies is a long run challenge. Mechanisms might involve differences in components that drive choice under uncertainty: access to information, mapping of information into subjective probabilities of outcomes and risk preferences Machina (1990), as well as constraints affecting capacity or ability to respond. For example, studies highlight the existing intersection of income and unequal access to information Norris (2001), differences in political preferences that may influence how information is processed Feldman and Johnston (2014), and attitudes towards risk Yesuf and Bluffstone (2009). Lower income households are also directly constrained in many ways, for example in the capacity to work from home, take paid or unpaid time off of work, and draw on savings to limit shopping trips to meet basic needs Orhun and Palazzolo (2019). Concentration of these households in denser residential areas may also complicate social distancing. With respect to environmental risk more broadly, environmental economists have emphasized the role of private defensive investments, whereby individuals invest in measures that reduce their exposure to environmental harms (e.g., air filters to reduce exposure to air pollution), and the extent to which such investments are limited by their income Deschênes et al. (2017).

While the data and analysis in this Brief do not yet allow for disentangling the causal roles of these various drivers, as a first step in this long run effort we document stark income-based differences in the response to recent state-level COVID-19 emergency declarations in the U.S.

## 2.2 Results

Fig. 2.1 shows the daily average of four mobility measures for weekdays in the U.S. from January 1 to April 21, 2020. These measures are averaged by income quintile at the smallest spatial unit available, either census tract (left column) or county (right column). Each panel shows a different daily mobility measures

Figure 2.1: Daily mean mobility measures in the U.S. from mobile devices for weekdays from January 1 to April 21 by quintiles of median income at the census-tract (left column) or county (right column) level. Thicker lines indicate the top and bottom quintile. Each panel shows a different measure of social distancing behavior. Data from SafeGraph, PlaceIQ and Google.



derived from mobile device location pings data: percentage of devices staying *completely at home* (top left); *device exposure* given by the average number of devices at all the locations visited by a device in a day (top right); *median distance traveled* outside the home, computed by taking the median distance traveled among the devices that left their home (bottom left); and percentage change in device presence at locations of *retail and recreation* relative to the 5-week period from January 3 to February 6.

All measures in Fig. 2.1 show an abrupt shift occurring in the month of March consistent with social distancing. They also show a clear pattern by income level: distancing responses range systematically from weakest for the bottom income quintile to strongest for the top income quintile. Notably, for the *completely at home* variable, which we view as the most appealing measure of social distancing, the income differential is reversed after March: individuals in the wealthiest census tracts shift from least to the most likely to completely stay at home, and vice versa for the poorest census tracts.

The *device exposure* measure provides a proxy capturing how often people are going to places combined with how crowded those places are. We found that the pre-pandemic income-mobility disparity—where high-income counties have substantially higher exposure—converges to parity under the pandemic. The *median distance traveled* income pattern diverges from the income-ranked relationship observed for other measures. This distance increases then decreases as income quintile increases, i.e., middle income quintile travelers typically move further in a day than the lowest and highest income quintile. This pattern holds before and after the arrival of the pandemic with the highest income quintile again showing the biggest change in behavior consistent with social distancing. The time series discussed above suggest that income differences are key determinants of mobile devices-based mobility measures. However, poorer and wealthier census tracts (or counties) are located in areas with different characteristics, where measures inducing social distancing may have been mandated at different times. To address both concerns in a multivariate framework, we use a panel regression analysis with an event-study design to estimate how social distancing behaviors are related to state emergency declarations, and how the response varies by income group (see Equation 2.1 in **Model**). The event-study model includes county fixed effect to control for the main effect of time-invariant county-specific factors such as income and population density, and also includes day fixed effects to control for temporal shocks common to all counties (e.g., federal policies). Further, as it is standard with event-study designs, we estimate the dynamic impact of the mandates by including lags and leads of the event date Bento et al. (2020). More specifically, we estimate a separate response for each of twenty days before and after an emergency declaration (normalized to occur at relative day 0), for each income quintile, allowing us to compare income-differentiated response to the state declaration. Coefficient estimates for days preceding each state's declaration inform us about potential pre-trends, i.e., changes in behavior ahead of the policy announcement. In a classical event-study analysis, substantial pre-trends are a source of concern, e.g., indicating pre-policy response.

Fig. 2.2 shows event-study estimates of the change in the four separate mobility measures (each panel) relative to the state emergency declaration date. Estimates for each measure (panel) are differentiated by county income quantile (lines within each panel). We find that pre-trends are absent from our preferred social distancing measure *completely at home* as well as *device exposure*. In contrast, for *median distance traveled* we find that individuals in wealthier counties show substantial behavior change before their state's declaration, while *retail and recreation* shows an early pre-trend that is then absent in the two weeks before the event. Post-event, however, for the majority of measures we estimate large and persistent social distancing behavior that is also strongly differentiated by income quintile. For example, for *completely home*, relative to the post-declaration average for all counties, the top income quintile response is essentially doubled while the bottom income quintile response is halved. Substantially more social distancing for the top income quintile is also apparent for *device exposure* and *retail and recreation*.

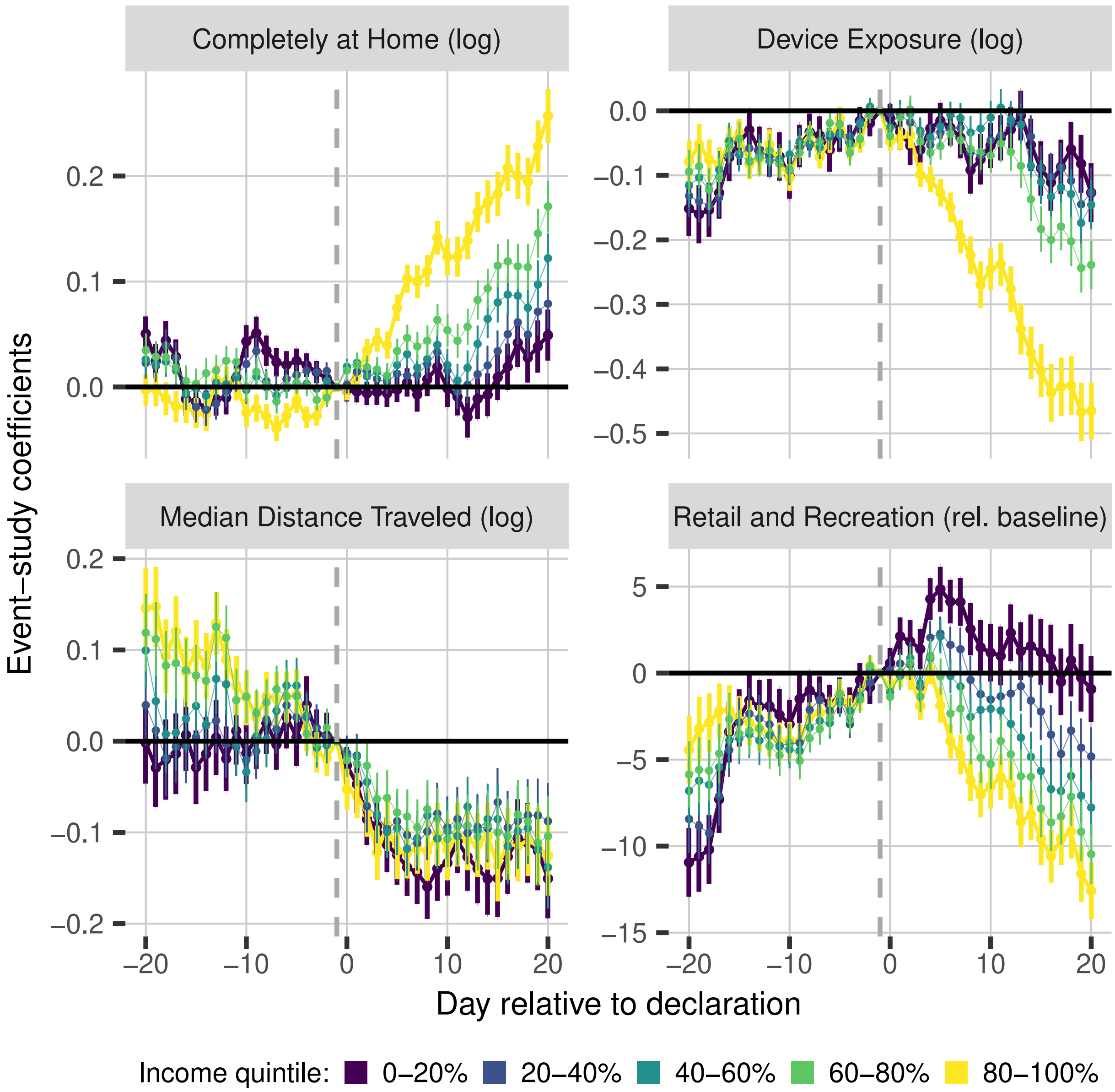
Explicitly identifying the relationship between social distancing and reductions in COVID-19 incidence is a key long run research need. However, the likely presence of two-way feedback between distancing and disease will require a creative and comprehensive approach for responsible causal inference. While state emergency declarations were typically the first major steps taken in most jurisdictions, states also took subsequent measures and thousands of counties and cities took steps at a finer spatial scale—a longer run exhaustive accounting of the policy drivers of distancing should account for these.

Overall, we show that social distancing following states' emergency declarations is substantial and strongly differentiated by county-level income. While these top-line results are consistent across distancing measures, other differences between them (e.g., income quantile convergence in *device exposure* but not in other measures) show the importance of considering multiple metrics. Our findings are in line with previous research showing an association between income and response to government prompts for disaster preparedness Beatty et al. (2019) as well as increased exposure of lower-income populations to environmental harms such as air pollution Downey and Hawkins (2008). The results highlight the urgent need for policy options to build capacity for social distancing—and other COVID-19 risk reduction measures—in lower income regions.

## Data sources

We assembled a longitudinal dataset of daily mobility measures and state-level emergency declarations for January-April, 2020. Daily mobility measures based on anonymized and aggregated mobile device data were obtained from Safegraph, Google and Place IQ. Safegraph data (*completely home*; *median distance traveled*) are provided at the census-block group level (period January 1 - April 21). PlaceIQ data was used

Figure 2.2: Event-study estimates of the change in a mobility measure (each panel) relative to state emergency declarations, differentiated by median income quintile at the census-tract (left column) or county (right column). Error bars represent 90% confidence intervals. State declaration occurs at day $k = 0$.



by Victor Couture and Williams (2020) to derive the *device exposure* variable at the county-level (period January 20 - April 21). Google Mobility data (*retail and recreation*) are provided at the county-level over the period February 15 - April 21 (expressed in changes relative to the 5-week period from January 3 to February 6, 2020). Of the three datasets, Safegraph data provides the best spatial coverage with almost all counties represented. Google mobility data was not consistently available for all days and counties due to anonymity constraints. More detail on specific measures is provided in **Data access**. U.S. states emergency declaration dates (date effective) were obtained from the National Association of Counties and updated with media reports. Finally, county and census-tract median-income quantiles were constructed using American Community Surveys data (2014-2018, 5-years pooled).

### 2.2.1 Model

We estimated the impact of a state's emergency declaration on mobility outcome $Y_{ct}$ using the event-study framework in Equation 2.1, where $Y_{cd}$ is the mobility measure in county $c$ on calendar day $d$. We index income quintiles by $q \in Q$ with $Q = \{1, 2, ..., 5\}$ and days relative to event by $k$. $D_{ckq}$ is a dummy variable equal to 1 when county $c$ belonging to income quintile $q$ is $k$ days away from being "treated" by the state declaration. Formally, $D_{ckq} = 1\{d - ED_c = k \cap q_c = q\}$, where $ED_c$ is the state emergency declaration day for county $c$. As is usual with event studies, we also included a single dummy for all relative days before our event window (over 20 days pre-declaration) denoted by $k = -21^{-}$, and another for all relative years after, $k = 21^{+}$ (over 20 days post-declaration).

$$Y_{cd} = \sum_{k=-21^{-}}^{k=21^{+}} \sum_{q \in Q} \gamma_{kq} \cdot D_{ckq} + \theta X_{cd} + \lambda_c + \lambda_d + \epsilon_{cd} \tag{2.1}$$

The dummy for $k = -1$ is omitted to serve as baseline. We further add as a time-varying control variable the cumulative number of COVID-19 infected cases in each county ($X_{cd}$). County-fixed effects ($\lambda_c$) control for unobserved static differences between counties, and day fixed effects ($\lambda_d$) control for national-level shocks. Finally, $\epsilon_{cd}$ is a county-day specific error term. Standard errors are clustered at the county level.

### 2.2.2 Additional robustness checks

We ran similar specifications using five different mobility measures provided by Safegraph (*full time work behavior*, *part time work behavior*, *median home share*, *median home dwell time*, and *delivery behavior*) and the PlaceIQ *device exposure* variable, both in levels and in logs, as well as the Google mobility variables (*retail and recreation*, *grocery and pharmacy*, and *workplaces*). Key results are consistent across all these alternative specifications. We alternatively included the cumulative number of known infected cases until the state declares an emergency, or excluded this variable altogether. This did not substantially affect our findings. We also conducted the analyses by income deciles instead of quintiles, which provided very similar results.

### 2.2.3 Data access

Mobility data are available from PlaceIQ via the Device Exposure variable derived by and available from Victor Couture and Williams (2020); Google Mobility Reports (available at https://www.google.com/covid19/mobility/); and Safegraph (freely provided upon request submitted at https://www.safegraph.com/covid-19-data-consortium). ACS data are provided on the American Community Survey website. State emergency declaration dates and code used to run the models are available on Github https://github.com/JoakimWeill/covid_mobility_income_PNAS.

We thank Safegraph, PlaceIQ and Google Mobility for making data publicly available and Arnon Erba for technical assistance. This research was supported by the California Breast Cancer Research Program of the University of California (UC), Grant Number R00RG2419.

## 2.3 References


Aizer, A. and J. Currie (2014). The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science 344*(6186), 856–861.

Beatty, T. K., J. P. Shimshack, and R. J. Volpe (2019). Disaster preparedness and disaster response: Evidence from sales of emergency supplies before and after hurricanes. *Journal of the Association of Environmental and Resource Economists 6*(4), 633–668.

Bento, A. I., T. Nguyen, C. Wing, F. Lozano-Rojas, Y.-Y. Ahn, and K. Simon (2020). Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. *Proceedings of the National Academy of Sciences of the United States of America*, 2–4.

Berchick, E. R., E. Hood, and J. C. Barnett (2019). Health insurance coverage in the United States: 2018. Current population reports. Washington DC: US Government Printing Office.

Bitler, M., H. Hoynes, and D. W. Schanzenbach (2020). Why the Safety Net Might Not Respond as Effectively to COVID-19 As It Should. *Milbank Quarterly Opinion*.

Centers for Disease Control and Prevention (2020). COVID-19 in racial and ethnic minority groups. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/racial-ethnic-minorities.html. Accessed: 2020-06-24.

Deschênes, O., M. Greenstone, and J. S. Shapiro (2017). Defensive Investments and the Demand for Air Quality : Evidence from the NOx Budget Program. *American Economic Review 107*(10), 2958–2989.

Downey, L. and B. Hawkins (2008). Race, Income, and Environmental Inequalities in the United States. *Sociological Perspectives 51*(4), 759–781.

Feldman, S. and C. Johnston (2014). Understanding the determinants of political ideology: Implications of structural complexity. *Political Psychology 35*(3), 337–358.

Machina, M. J. (1990). Choice under uncertainty: Problems solved and unsolved. in *National Research Council (US) Steering Committee on Valuing Health Risks, Costs, and Benefits for Environmental Decisions*. (National Academies Press).

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide (Communication, Society and Politics).* Cambridge University Press.

Orhun, A. Y. and M. Palazzolo (2019). Frugality Is Hard to Afford. *Journal of Marketing Research 56*(1), 1–17.

Springborn, M., G. Chowell, M. MacLachlan, and E. P. Fenichel (2015). Accounting for behavioral responses during a flu epidemic using home television viewing. *BMC infectious diseases 15*(1), 21.

Victor Couture, Jonathan Dingel, A. G. J. H. and K. Williams (2020). Exposure indices derived from placeiq movement data. https://github.com/COVIDExposureIndices/COVIDExposureIndices. Accessed: 2020-05-01.

Yesuf, M. and R. A. Bluffstone (2009). Poverty, risk aversion, and path dependence in low-income countries: Experimental evidence from Ethiopia. *American Journal of Agricultural Economics 91*(4), 1022–1037.

# Chapter 3

# Researchers' Degrees-of-Freedom: Evidence From the Pandemic Policy Evaluations

**Published as Researchers' Degrees-of-Flexibility and the Credibility of Difference-in-Differences Estimates: Evidence From the Pandemic Policy Evaluations. Joakim A. Weill, Matthieu Stigler, Olivier Deschenes, and Michael R. Springborn. National Bureau of Economic Research, 2021.**

## 3.1 Introduction

The COVID-19 crisis has brought unprecedented responses by national and sub-national authorities. At the center of this public action has been mobility restrictions and lockdown orders aimed at reducing the risk of virus transmission through constrained social contacts. More than a year after their initial implementations, and as subsequent waves of COVID-19 cases surged in many countries around the world, such measures continue to be reintroduced and their efficacy debated. Therefore a key question is whether, and to what degree, these mobility-reducing policies cause reductions in measured mobility.

Due to the urgency of this question, and the near-real time availability of data, the scientific literature focusing on the effect of what we refer to as 'mobility-restricting policies' (MRPs)[1] on mobility grew at near exponential rate: as of March 2021, a Google Scholar search of "mobility" and terms related to MRPs implemented in response to the COVID-19 pandemic ("shelter-in-place", "stay-at-home", or "lockdown") returned over 23,000 results. A large number of these studies focus on estimating the impacts of mobility-restricting policies on measured mobility with the level of analysis varying from city to worldwide. In particular, due to the sharp temporal adoption of MRPs, many of these studies leverage a difference-in-differences (DD) based design. This previous literature does not reach a consensus; results range from finding that "there is little evidence... that stay-at-home mandates induced distancing" (Gupta et al., 2020) to "clear effects of stay-at-home orders on social distancing" (Allcott et al., 2020). At the same time, several studies highlight existence of pre-trends, often interpreted as indicating that voluntary mobility reductions were an important factor (e.g. Gupta et al. 2020; Andersen 2020; Abouk and Heydari 2021).

In this paper, we re-examine the evidence from this early literature and assess the impact of mobility restrictions on county-level mobility in the United States, leveraging MRPs uniquely gathered across state, county, and city levels. First, we document and correct important gaps in the source data on mobility restriction policies used in the previous literature. For example, one commonly used data set indicated that around one third of U.S. counties enacted an emergency declaration by the end of March 2020 (NACo, 2020); our data gathering showed that the true number was twice as large. Using the corrected data, we consider a broad set of five MRPs: State emergency declarations, State shelter-in-place orders, State earliest restriction or closure, County emergency declarations, and County shelter-in-place orders. We supplement these MRPs data with an extensive set of 20 daily mobility indicators derived from anonymized mobile device ping information provided by Safegraph, Google Mobility, PlaceIQ, and Cuebiq.

Second, we estimate the impact of MRPs on observed mobility using standard two-way fixed effects and event-study methods, following the approach used in the previous literature. We examine a wide range of specifications that were used in the previous literature (although in distinct individual papers), focusing on linear models, log-linear models, seven-day moving averages, and first-differences. We further supplement these analyses by considering all mobility outcomes observed at the county-day level, and also estimate the impact of MRPs individually and in combination. The later is especially important since MRPs are almost always overlapping (e.g., counties with a county shelter-in-place order may also have a county emergency declaration or a state shelter-in-place order) and temporally dependent (e.g., emergency declarations typically precede shelter-in-place orders).

Most importantly, the wide range of specifications we consider allows us to assess whether the estimates

[1] These policies are also described as "non-pharmaceutical interventions (NPIs)" in the literature

of the impact of MRPs on mobility in the prior literature paint a representative picture of the evidence. In particular, we emphasize the wide range of researchers' degrees of flexibility (or freedom) that exist over many modelling and specification choices: the outcome variable (mobility measure) and its functional form in the regression model, covariates included, estimator, and the spatial and temporal unit of analysis (e.g., state or county, daily or weekly). Rather than focusing on a "clean" and robust set of results in support of a consistent narrative, we report estimates along many dimensions of the degree-of-flexibility space, while highlighting contradictory findings, failing robustness tests, and ambiguous results.

At the county scale, we find that some conclusions drawn about the efficacy of mobility restricting policies in the previous literature using two-way fixed effect regressions are not robust to seemingly minor variations in the estimating model, casting doubts on the internal validity of the prior estimates. Depending on the mobility outcome and policy considered, several estimates have the "wrong" sign, where the MRP is found to increase mobility. For a given mobility measure, several estimated effects switch sign, but remain statistically significant, when using the log of the outcome variable instead of the level. Most alarmingly, the standard robustness tests based on the sequential addition of covariates fail to provide warning signs of specification sensitivity. The fact that robustness on covariates does not imply robustness on transformations of the outcomes illustrates that focusing on "one-dimensional" robustness checks can be misleading.

Third, we find that event-study estimates in this context are plagued with pre-trends, which have been interpreted by previous studies as evidence of voluntary social distancing in anticipation of the policy announcement. These issues are most pronounced using the standard event-study estimators. We apply recent heterogeneity-robust DD estimators and find that they partially mitigate these issues, but only for certain outcome variables and functional forms. Overall, we document that at the county-day level, the existing tools for difference-in-differences analysis —both standard and recently developed methods—are insufficient for making rigorous and specific conclusions regarding the causal impact of MRPs on mobility. To be clear, we do not assert that MRPs were ineffective; strong patterns in the data are highly suggestive of a substantial impact on mobility. Rather, our conclusion is that DD-based evaluation methods are not yet a sufficient match for the complexity presented by multiple overlapping policies and feedback processes to the behavioral outcome of interest at a fine-scaled spatio-temporal level (i.e., county-by-day). The ultimate concern of broader policy interest is not just whether MRPs reduced mobility but also whether actual infections and deaths were curtailed. However, this latter step requires rigorous and robust treatment of the first step, which we show is much less straightforward than initially expected.

Besides our direct contribution to the literature on the efficacy of mobility-restricting policies, our paper brings to light issues related to the meta-scientific literature on researchers' degrees of flexibility and the credibility of economics research (see Christensen and Miguel (2018) and Kasy (2021) for recent overviews).

In any empirical study, researchers are confronted with a multitude of decisions on how to handle the data and conduct the analyses, a *garden of forking paths* in Gelman and Loken (2013)'s terms. The resulting findings are conditional on the specific path of decisions made along the way, leaving open the question of whether different paths would have lead to very different conclusions.

The sensitivity of empirical results to researcher degrees of flexibility have been highlighted in several replication studies (Aiken et al., 2015; Foote and Goetz, 2008; Clemens and Hunt, 2019), as well as in more exhaustive "many-analysts" experiments, where multiple research teams are asked to analyse the same data set (see e.g., Silberzahn and many authors (2018); Botvinik-Nezer and many authors (2020); Huntington-Klein et al. (2021)). These studies are based on artificial research exercises, where each researcher's goal is not to write a full-length article but instead to produce an isolated analysis and set of findings. One might hope that in the process of actual research, major instability of the results over alternative researcher's decisions would be discovered and highlighted, either in different papers or within the "robustness" section of a given paper. The recent COVID-19 literature provides a natural experiment to assess whether this is actually the case: Due to the importance and urgency of the COVID crisis, a remarkable number of studies focusing on the same research question were released and/or published in a very short period of time. This represents a unique opportunity to investigate in a real setting both the influence of researchers' degrees of flexibility on the study's conclusions; and whether the sensitivity of the results is reflected in each paper or across the combined literature.

By estimating MRP impacts on mobility along all main dimensions of the researchers degrees of flexibility in our setting, we can point to several concerns that have not been fully addressed in the previous literature. We find that apparently innocuous researcher's decisions such as modelling the dependent variable with the log transformation or leaving it in levels can have a large impact on the results, even impacting the signs and statistical significance of the estimated effects. Further, we find that for almost every *well behaved* result obtained, we can find an alternative and equally reasonable specification which gives opposite sign or null results. These issues are mitigated but not fully resolved by the use of recent heterogeneity-robust DD estimators (e.g., Callaway and Sant'Anna 2020). Finally we find pervasive pre-trends in the data for virtually all MRP and mobility outcomes, which can only be satisfactorily addressed by adopting a particular specification.

We do not take a stance as to why the previous literature largely focuses on highlighting the 'clear' impacts of MRPs on mobility. This could be due to random factors, the supply side of research production (e.g., selective reporting of results in agreement with the researchers' priors, selective submission of results that provide a consistent story, biased sampling of the specifications space when exploring possible models), or the demand side (anticipated need by researchers to provide 'clean' results and story as a necessary step for publication). Our own previously published research in this area highlighted significant impacts of emergency declarations (one prominent example of MRP) on mobility outcomes, though also documented large pre-trends in event-studies (Weill et al., 2020). Furthermore, the econometric research focused on issues with two-way fixed effects specifications developed extremely rapidly, and in parallel with the applied COVID-19 policy research. One paper cautioned about identification issues with difference-in-differences methods in the context of COVID-19 policies (Goodman-Bacon and Marcus, 2020), but its recommendations were largely ignored. Many of the shortcomings linked with the use of two-way fixed effects estimators under staggered adoption designs have only recently become apparent. While this paper does not attempt to identify the specific sources of violations of the identifying assumptions, we show the existence of warning signs in the estimates that might lead a researcher to reconsider the validity of the research design, while simultaneously showing how researcher's degrees-of-flexibility might obfuscate (willingly or unwillingly) the magnitude of these issues.

Concerns about selective reporting of results, p-hacking, and pre-testing are not new. Leamer (1978) recommended more than 40 years ago that researchers should report the results of all estimations that they tried, as opposed to focusing on a few chosen ones. Similarly, in the context of event-studies, where the causal interpretation of a policy effect relies on the absence of pre-trends, recent research shows that focusing on pre-trend tests that "work" can induce substantial bias in the reported estimates (Roth, 2019). Unfortunately, the recommendation made by Leamer is far from being a standard practice today. When featuring alternative specifications to the main model, published studies tend to report "robustness tests" that are based on the inclusion of alternative covariates in the regression model. These reported robustness tests also generally agree with the main findings of the research. As we show in this paper, failure to consider robustness analysis along all relevant dimensions (outcomes under study, functional form, covariates, and estimator) can lead one to wrongly infer that the estimated effect is stable.

The remainder of this paper proceeds as follow. First, we summarize a selected set of existing empirical studies of the impact of MRPs on mobility (directly below). We then present our analysis, beginning with the data (Section 2) and results of the standard two-way fixed effects models (Section 3). Next, we explore robustness tests (Section 4) and contradictory results from event studies (Section 5). Finally, we round out the analysis with results from heterogeneity-robust estimators (Section 6).

### 3.1.1 Review of existing empirical evidence

Table **??** summarizes notable studies that address similar questions using difference-in-differences designs include a set of working papers (Gupta et al., 2020; Andersen, 2020; Elenev et al., 2021) and peer-reviewed articles (Painter and Qiu, 2021; Villas-Boas et al., 2020; Allcott et al., 2020; Abouk and Heydari, 2021; Dave et al., 2021; Goolsbee and Syverson, 2021). While the shared methodology might suggest a relatively uniform approach, examination of the literature indicates the absence of a consensus on many empirical design choices. For example, each analysis uses a different transformation of the key outcome/dependent variable (mobility metric), including levels (Gupta et al.); natural log (Allcott et al.; Goolsbee and Syverson; Elenev et al.); the change relative to a fixed pre-COVID baseline in level or log (Villas-Boas et al.; Dave et al.); the daily change relative to a daily 2019 counterfactual (Andersen); and daily first differences (Painter and Qiu; Abouk and Heydari). Four of these articles use weighted regression population weights, one used weights based on the number of visits to stores in January (Goolsbee and Syverson), another via a synthetic control method (Villas-Boas et al.), while the rest do not. Mobility data are typically drawn from a single source, with Safegraph being the most common. Most studies focus only on mobility response measured at the state level. Exceptions include Allcott et al. who focus on the county level but combines county and state policies into one metric. Gupta et al. also focuses on the county level and considers state and county policies, albeit not with full event study regressions where the different policy effects are estimated simultaneously. These analyses typically rely on existing databases of reported policies (e.g., such as those collected by the Kaiser Family Foundation or the New York Times) and none (to our knowledge) gather substantial new data on which counties or cities enacted local policies.

Table 3.1: Selected Studies Analyzing the Impact of MRPs on Mobility Outcomes

| Article | Dependent variable transformation | Population weighted | Mobility data source | MRPs | Units |
|---|---|---|---|---|---|
| Gupta et al., 2020 | no | no | PlaceIQ, Safegaph, Google, Apple | county and state | county and state |
| Painter and Qiu, 2020 | first-difference | no | Safegraph | state | county |
| Andersen, 2020 | deviation from 2019 counterfactual; log for visits | yes | Safegraph | state | state (event studies) |
| Villas-Boas et al., 2020 | %-deviation from the 2/10-3/8 period | no (weight with SCM) | Unacast, Google | state | state |
| Allcott et al., 2020 | log | yes | Safegraph | county | county |
| Abouk and Heydari, 2021 | first-difference | yes (TWFE) | Google | state | state |
| Dave et al., 2021 | deviation from the 2/6-2/12 period | yes | Safegraph | state | state |
| Goolsbee and Syverson, 2021 | log | yes (various) | Safegraph | county and state | county |
| Elenev et al., 2021 | raw shares; log for visits | yes and no | Sagegraph | county/state merged | county |

*Notes:* SCM = synthetic control method, TWFE = two-way fixed effects.

Issues with pre-trends are apparent and discussed in several of these articles. Many observe that variation in mobility is not fully or even largely explained by MRPs. For example, Allcott et al. find that "the magnitude of policy effects is modest, and most social distancing is driven by voluntary responses". Similarly Abouk and Heydari conclude that "(a)lthough evidence for reduced social contact in the United States is strong...

people in most states had already started to reduce the time they spent outside their homes before any NPI (non-pharmaceutical intervention) was implemented", and Goolsbee and Syverson concludes that "[...] the vast majority of this drop is due to individuals' voluntary decisions to disengage from commerce rather than government-imposed restrictions on activity." Only one article focuses on the prospect of time varying treatment effects in its main results: Andersen observes that "it is possible that difference-in-differences and event study estimates are biased by comparisons between early and late treated units". Within this literature we could find no discussion focused on estimates that have an unexpected sign or that switch sign across plausible alternative specifications, as we do below. Taking the previous literature as a starting point, we have designed our paper to fully explore the implications of empirical design choices (data sources, estimator, functional form choice, policy under consideration, etc) on the estimated impact of mobility-restricting policies on mobility. Finally, one article explicitly investigates the role of spillover effects between nearby counties within the framework of two-way fixed effects models (Elenev et al.).[2]

## 3.2 Data Sources and Preliminary Analysis

Our empirical analysis of the impact of mobility restricting policies on mobility outcomes is conducted with a comprehensive set of data files on mobility outcomes collected from mobile device signals, combined with extensive MRP data assembled at the city, county and state level. This section describes the data sources, defines the primary outcome and control variables, and then presents some summary statistics.

### 3.2.1 Mobility Measures

We assembled a daily data set of 20 mobility measures at the county level collected during the first-wave of the COVID-19 pandemic in the United States. We restrict our analysis to start in January 2020 and to end on April 21st, 2020. This corresponds to the time span between the earliest date the mobility data are available, and the date where some jurisdictions began to lift distancing orders on April 21st, 2020. Our mobility measures are all based on anonymized and aggregated mobile device signal ("pings") data and come from four different sources. The variables are listed in Table 3.2 along with their source and summary statistics. Additional information on each of the variables appears in Table G.1 in the Appendix. Safegraph data are provided at the census-block group level, which we aggregate to the county level. PlaceIQ data were used by Couture et al. (2020) to derive the *Device Exposure* variable at the county-level. Google Mobility

[2]Based on leading epidemiological models, Callaway and Li (2021) argue that identification approaches based on unconfoundedness are more likely to outperform difference-in-differences models, and propose a more structural approach estimating the impacts of mobility-restricting policies. Chernozhukov et al. (2021) also use a variant of a SIR epidemiological framework to motivate a structural model for the potential outcomes. These more structural approaches are outside the scope of this paper.

data are provided at the county-level over the period February 15 - April 21 (expressed in changes relative to the 5-week period from January 3 to February 6, 2020). Cuebiq data are provided at the county level.

Table 3.2: Summary Statistics and Sources for Mobility Outcomes

| Source | Variable | Re-signed | N | Counties | Mean |
|---|---|---|---|---|---|
| Safegraph (SG) | Average Time Not Home | yes | 285,405 | 3069 | 284.55 |
| | Away at least 3 hours (share) | yes | 285,405 | 3069 | 0.15 |
| | Completely Home (share) | no | 285,405 | 3069 | 0.27 |
| | Full Time Work (share) | yes | 285,405 | 3069 | 0.06 |
| | Median Distance | yes | 285,405 | 3069 | 14104.85 |
| | Median Home Dwell Time | no | 285,405 | 3069 | 616.82 |
| | Median Home Share (share) | no | 285,405 | 3069 | 0.74 |
| | Median Not Home Dwell Time | yes | 285,405 | 3069 | 139.43 |
| | Part Time Work (share) | yes | 285,405 | 3069 | 0.09 |
| PlaceIQ (PQ) | Device Exposure | yes | 183,117 | 1969 | 89.82 |
| Google (GM) | Grocery & pharmacy | yes | 134,843 | 2425 | 1.19 |
| | Parks | yes | 42,445 | 944 | 8.84 |
| | Residential | no | 77,185 | 1526 | 7.95 |
| | Retail & Recreation | yes | 139,935 | 2517 | -12.21 |
| | Transit stations | yes | 64,358 | 1105 | -13.11 |
| | Workplaces | yes | 164,680 | 2722 | -18.87 |
| Cuebiq (CQ) | Mobility Index | yes | 285,413 | 3069 | 3.65 |
| | Staying Around Home (share) | no | 285,413 | 3069 | 0.28 |
| | Staying Around Neighborhood | no | 285,413 | 3069 | 0.37 |
| | Traveling more than 10 miles | yes | 285,413 | 3069 | 0.34 |

Notes: Table 2 reports summary statistics for the 20 mobility indicator variables analyzed in this paper. The number of observations corresponds to the number of counties in which a mobility indicator is observed multiplied the number of days it is observed. The "re-signed" column indicates if a mobility indicator will be re-signed in the analysis below so that increase in the variable can be interpreted as increase in social distancing. In the tables and figures below re-signed variables will be identified with an asterisk next to their name.

The summary statistics in Table 3.2 include the number of county-day observations ($N$) for which each of the mobility measures are available between January and April 2020. Differences in the temporal and spatial coverage of these measures are evident, with Safegraph and Cuebiq providing the best spatial coverage with almost all counties represented (over 285,000 observations, amounting to 3069 counties and 93 days). Google mobility data are not consistently available for all days and counties due to Google's anonymity constraints. Many of these variables are designed to capture the amount of time spent at or away from home by individuals,[3] other variables correspond to specific activities outside the home,[4] while some other variables can be interpreted as mobility and social mixing indices.[5]

[3] *Staying Around Home (%) Average Time Not Home*, *Median Not Home Dwell Time*, *Completely Home (%)*, *Median Home Dwell Time*

[4] *Grocery and Pharmacy Retail and Recreation Workplaces*, *Full Time Work (%)*

[5] *Device Exposure*, *Median Distance*, *Mobility Index*

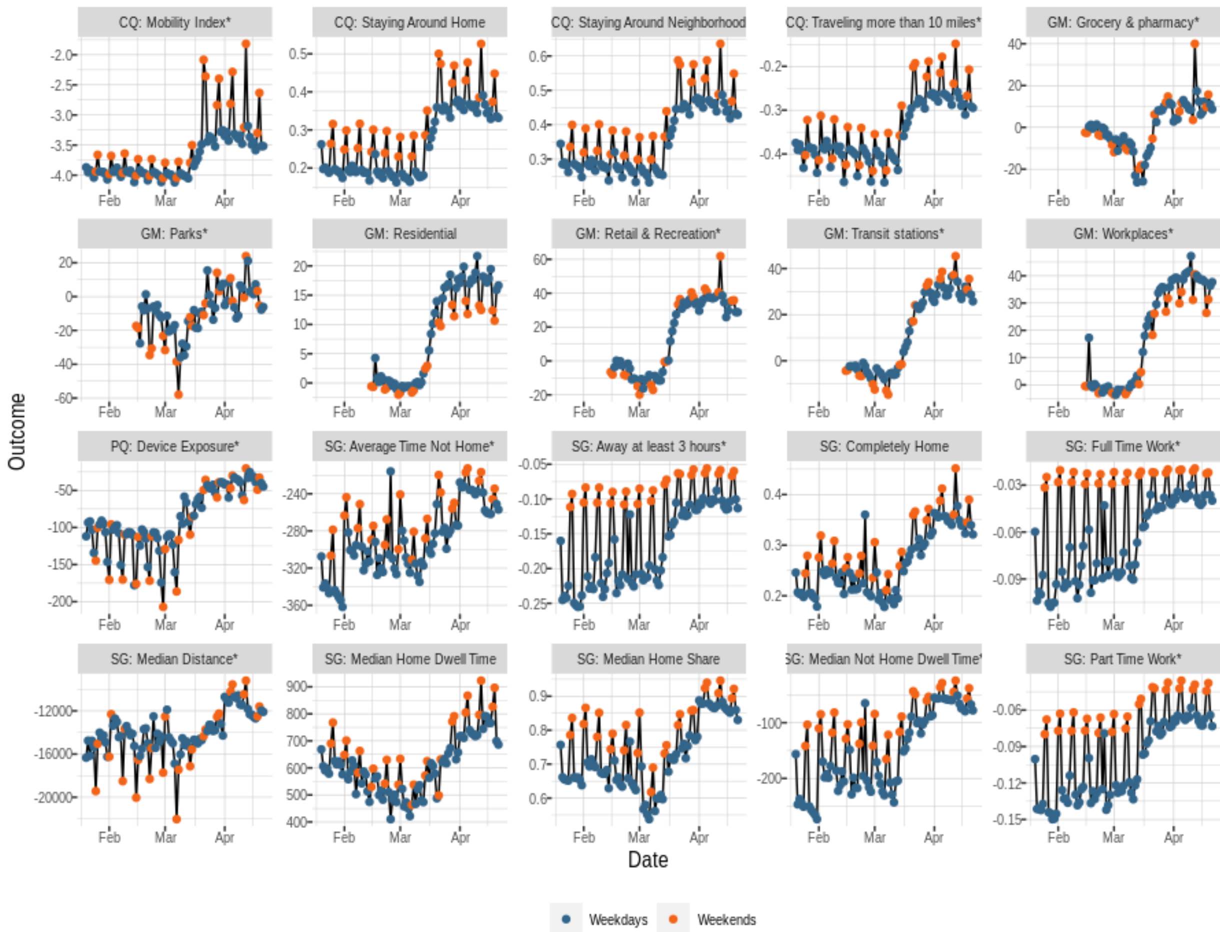
In the fourth column we report the sample mean of the various mobility measures. For example, the Safegraph variable *Completely Home* indicates that 27% of sampled devices spend the day completely at home, while the *Median Distance* travelled outside the home is 10,683 meters on average between January 20 to April 21. Similar to the information from Safegraph, Cuebiq data indicates that 28% of device users are *Staying Around Home*. The Google Mobility data shows both positive and negative averages, as each of its mobility indicators are normalized relative to same day of the week between January 3 to February 6, 2020, before the effects of the COVID-19 pandemic were fully realized in the U.S. Interestingly, within these relative Google measures we see that *Parks* and *Residential* time have positive averages, while *Retail & Recreation*, *Transit stations*, and *Workplaces* all have negative averages, consistent with changes in mobility behavior to increase social distancing.

Figure 3.1 plots the mobility and activity patterns in the anonymized "pings" data, with weekdays shown by red dots and weekend days shown with blue dots. **To ease comparisons between outcome variables, for the remainder of the paper we analyze "re-signed" mobility outcomes so that increase in the variable can be interpreted as increase in social distancing**. For example, *Traveling more than 10 miles* was multiplied by minus one: on the first row and fourth column of Figure 3.1, the variable indicates an initial average level of about -0.4 in January (meaning that on average at the county level, 40% of people moved further than 10 miles per day), and then increases quickly to -0.3 in April. When working with the log of the outcomes, we first take the log and then multiply by minus one. To keep track of this convention, variables that were re-signed are highlighted with an asterisk next to their name. Several notable features emerge from the simple analysis in Figure 3.1: First and foremost, there is an important reduction in mobility and social activity that occurs around the middle of March 2020, when the majority of the U.S. population was under some form of MRP. This reduction in mobility is evident for all of the outcome variables. For example, the Cuebiq *Staying Around Home* nearly doubles, and the Place IQ *Device Exposure* is halved. Second, the implied mobility by these indicators appears to be reduced more strongly during weekdays, consistent with the large increase in unemployment claims and in work from home observed at that time. Finally, for most of the mobility indicators, there is no reversion to pre-mid-March mobility and activity levels as of the of April 2020.

In addition to Figure 3.1, Figure H.1 in the appendix presents Pearson correlation coefficients between these different outcome variables used to measure mobility. Except for Google Mobility *Parks* and Safegraph *Median Distance*, all outcome variables are strongly correlated, with correlation coefficients well above 0.5. However, the analysis in Sections 3.4 and 3.5 will show that the choice of mobility outcome variable has an important impact on the estimated effect of MRPs on mobility.

Table E.1 in the appendix reports the sample averages of the control variables used in the empirical

Figure 3.1: Trends in Daily Mobility Outcomes, January-April 2020




Notes: Some mobility outcomes are resigned (indicated by a star next to the variable name) so that increases in a given outcome are interpreted as increases in social distancing. All mobility outcomes are in levels, except for Google Mobility which is only available in % change relative to a baseline.

analysis below. In our sample of 3069 counties, there are on average 2.14 confirmed cases of COVID-19 per day, and 27 % of counties-days observations have at least one confirmed case of COVID-19 (8% experienced at least one death attributed to COVID-19). Naturally, all these indicators of pandemic severity have means close to zero before the Mar 11, and much larger means afterwards. The remaining rows of Table E.1 report the average daily temperature (in °F), precipitation and snow (both in hundreds of inches).

### 3.2.2 Mobility Reducing Policies

We obtained data on the date of declaration of COVID-19 mobility reducing policies (MRPs) in the U.S. from February-April at the state, county and city level. For all three levels, our policies of interest include emergency declarations (EDs) and shelter-in-place orders (SIPOs). For state-level policies we used data provided by Fullman et al. (2020). For counties we used data collated by the National Association of

Counties (NACo, 2020) as a starting point. The NACo data constitute an incomplete set of county-level policies implemented. No pre-existing data set was available for cities. To fill in these gaps in county and city MRPs, we conducted a manual search for the dates of either type of declaration for (1) all U.S. counties lacking either a ED or SIPO in the NACo data, as well as (2) cities with a population of 50,000 or greater.[6] When existent, these dates were typically found in media reports, government websites or government meeting minutes. Overall, across 3,069 U.S. counties, the original NACo data set included 989 EDs and 147 SIPOs; these counts were slightly more than doubled (through manual search) in our final data set, to 1954 EDs and 334 SIPOs. Across the 783 U.S. cities with a population of 50,000 or greater, we identified 633 with EDs and 122 with SIPOs.

Governmental responses were not limited to EDs and SIPOs. For example, some states and counties issued separate orders for restrictions on restaurants, gatherings, schools and non-essential businesses. For consistency and tractability, given our coverage of three levels of government, we focus mainly on EDs and SIPOs. However, at the state level we also consider the "earliest restriction or closure" (ERC) as provided by Fullman et al. (2020), which takes the earliest state policy that either closed schools, restaurants, bars or restricted gatherings.

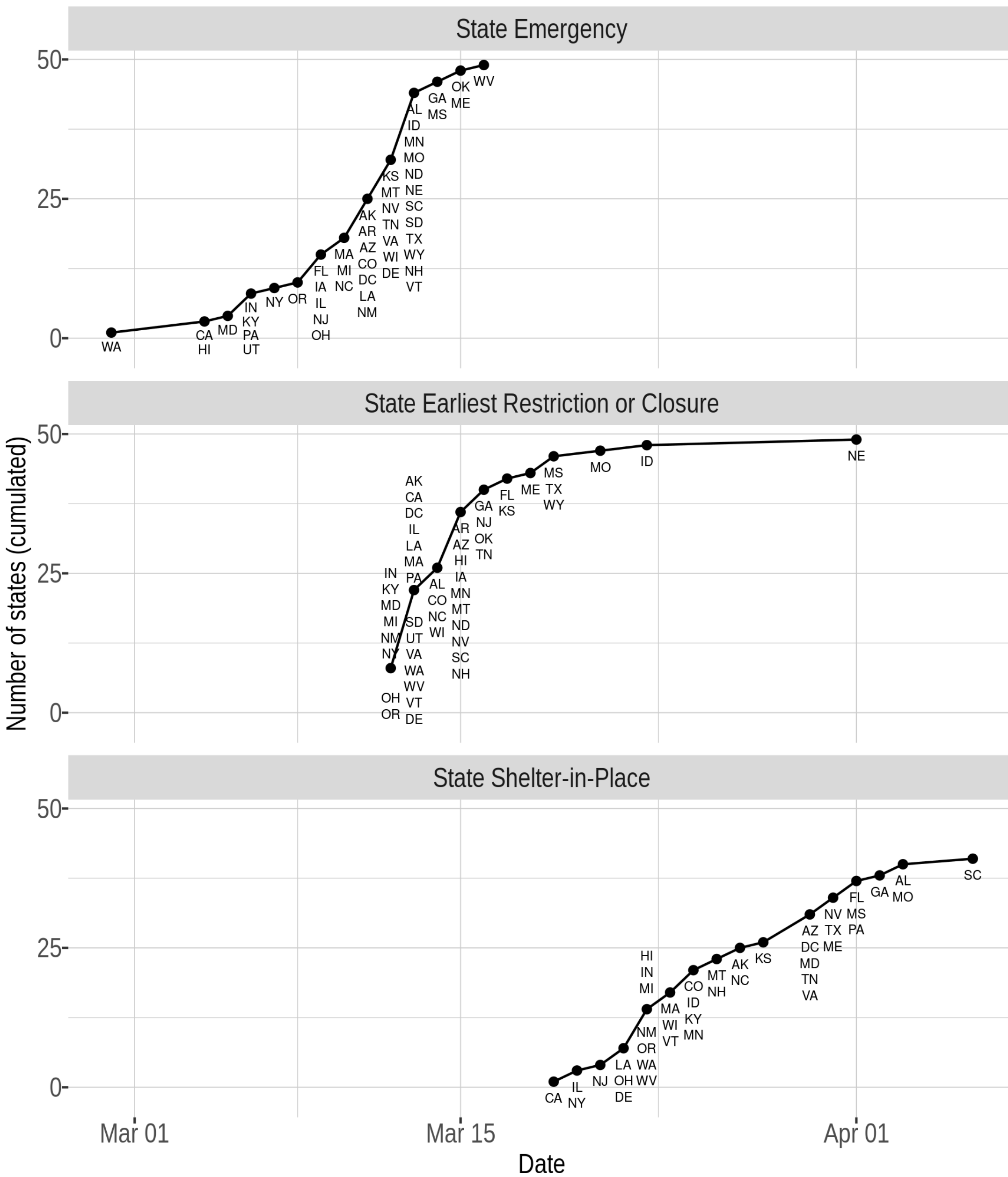
Figure 3.2 shows the cumulative adoption of state-level policies as a function of date, by policy and state (as indicated by state abbreviation). For example, Washington state was the first to declare an ED, followed by California and Hawaii. It is also evident that there is an ordering in MRP policy adoption, with the ordering starting with EDs, followed by ERCs, and then SIPOs. Notably, all states declared an ED and an ERC at some point before April 1st, but not all states imposed a SIPO.

Figures E.1 and E.2 in the appendix show the timing of county-level EDs and SIPOs adoptions, respectively. The main finding here is that county-level EDs are much more common than SIPOs. We also observe that EDs are less common in the less-populated areas of the central U.S., as well as the far northeastern U.S. This is not surprising as these regions were not heavily impacted by COVID-19 before April 2020.

Figure 3.3 displays the relative timing of the adoption of state and county policies. For each of the MRPs we consider, we compare pairs of policies (say "A" and "B") and reports the number of counties where policy A was adopted before policy B (blue line) and the number of counties where policy A was adopted after policy B (brown line). In addition, we report the average relative time (in days) between the adoption of policies A and B. For example, the top panel for state ED shows that 3014 counties (out of 3069 in the full sample) were in a state that adopted an ERC later than ED (with an average gap of 4 days), while 55 counties saw the opposite ordering. Looking at the comparison between state ED and the four other MRPs,

[6]We used population in 2019 as estimated by the U.S. Census Bureau (2020). Where county and city governments have been consolidated into one jurisdiction, we treat the entity as a county in our data set.

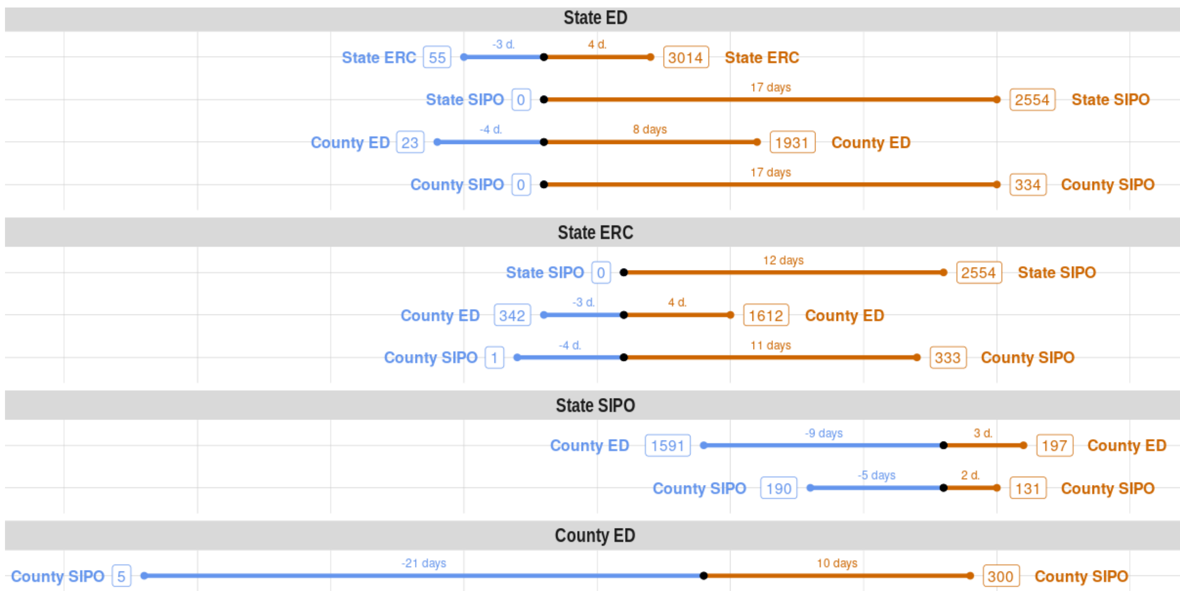
Figure 3.2: Cumulative Number of States Adopting a Given Mobility Restricting Policy by Date




Notes: Cumulative number of state-level MRP adoption by date. New adoptions on a given date are represented with state abbreviations.

it is evident that state EDs were generally adopted first, typically with long gaps before the adoption of a follow-up policy (e.g., 17 days for state SIPO and county SIPO). A similar pattern is observed comparing state ERC (typically the second MRP policy adopted after state ED) to the other MRPs. For example, all state SIPOs were adopted after the state ERC (12 days later on average), and all but one of the county SIPOs were put in place after the state ERC. Finally, county ED measures tend to be introduced before state SIPOs, while the relative timing of state and county SIPO adoptions tends to be balanced. Like the pattern for state-level policies, county SIPOs are introduced later that county ED measures, on average.

Figure 3.3: Pairwise Comparisons of the Relative Timing of Mobility Restricting Policy Adoption



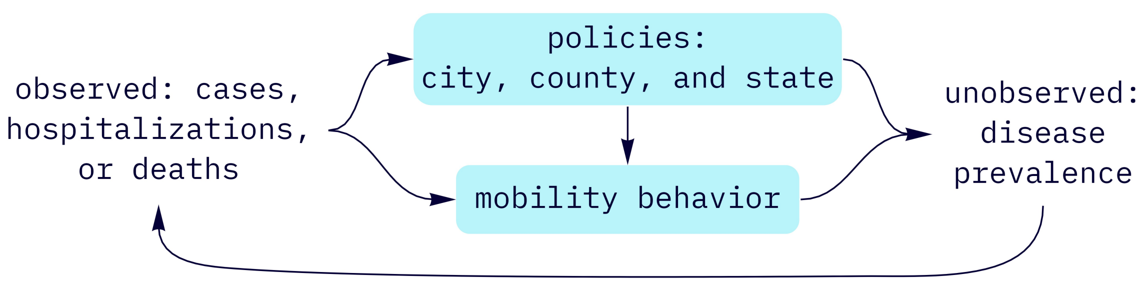

Notes: For each of the four MRPs listed in bold, the figure shows the average relative adoption time (reported in days) and the number of counties (reported in the boxes). Adoptions that come before are shown in blue and those that come after in brown. For example, in the top panel for state ED, we report that of all counties that ever adopted county ED, 1931 counties adopted state ED before county ED (with an average gap of 8 days), while 23 counties adopted county ED before state ED (with an average gap of 4 days).

There are three main conclusions from this section that have important implications for the empirical analysis below. First, considering state-level policies, all states implemented some form of MRP (although not all implemented the more restrictive SIPO), while at the county-level, 62 % of counties (representing 79 % of the U.S. population) implemented one or more form of MRP. Second, a substantial share of the U.S. population was subjected to multiple MRPs simultaneously (e.g., state ED, county ED, and county SIPO). Third, states and counties adopted MRPs in a staggered manner, with early adopters implementing the policies several weeks before later adopters. Further, there is typically a natural ordering of policy adoption, for example, state/county SIPOs always follows state/county EDs. These features of the policy response to the first wave of the COVID-19 pandemic complicate the estimation of the causal effect of various MRPs on mobility. Different jurisdictions often adopted a wide range of different policies, but not all of them: as a result, one might wish to estimate which of these MRPs work best, which motivates including them simultaneously in the analysis.

## 3.3 Policy impact estimation framework

Figure 3.4 represents our conceptual causal chain. We focus here on the link between MRPs and mobility behavior, and highlight two challenges that impede straightforward estimation of this relationship. First, rather than a single event involving one policy at one spatial scale, the early response to the COVID-19 pandemic included multiple policies at multiple scales. Furthermore, the vast majority of these policies appeared in a compressed window of two months, limiting the variation in "treatment timing" to inform the impact. Second, the conceptual causal chain we posit in Figure 3.4 raises the prospect of endogeneity. Our focus is estimating the impact of MRPs on mobility behavior. These policies by cities, states and counties are clearly a response to perceived risk as informed by COVID-19 cases, hospitalizations and deaths. However, we might also expect individuals to respond directly to these perceived risks, in addition to policies. Finally, these policies and changes in mobility behavior are expected to influence disease prevalence and hence future perceived risk as further disease outcomes are observed. In future work, it may be that a structural approach to modeling this system would provide further traction in disentangling these effects.

Figure 3.4: Conceptual causal chain. This article focuses on the relationship between policies and mobility behavior (shaded).




### 3.3.1 Empirical model specification

As a starting point for our model specification, we follow the previous literature and aim to estimate the impact of mobility reduction policies (MRPs), enacted by different jurisdictions (states and counties) at different points in time, using a linear two-way fixed effect estimator (TWFE):

$$Y_{it} = \alpha_i + \delta_t + \sum_{p=1}^{P} \beta_p D_{ipt} + X_{it}'\gamma + \epsilon_{it} \tag{3.1}$$

where $i$ denotes county, $t$ represents the day (from January 20 - April 21, 2020), and $p$ is the index for the five policies we analyze. $Y_{it}$ is one of the 20 different mobility and activity outcome variables we described in Tables 3.2 and Appendix Table G.1. $D_{ipt}$ is a dummy variable equal to 1 when county $i$ is "treated" by a given MRP policy $p$ within the set $P$: State ED (Emergency Declaration), State ERC (Earliest Restriction and Closure), State SIPO (Shelter in Place Order), County ED, and County SIPO on date $t$. To begin we follow the previous literature's practice of analyzing the impact of MRPs individually (i.e., by including them one at the time in the model), but we will also report estimates of $\beta_p$ when all the impact of all policies are estimated simultaneously.

The vector of control variables $X_{it}$ are observable predictors of mobility and activity, which include daily precipitation, snowfall and mean temperature. In some specifications we also include binary indicators for whether the first COVID-19 case and, separately, the first COVID-19 mortality has been recorded in a county.[7] Including these variables in the models have important implications in terms of the required identifying assumptions. On the one hand, such controls may help alleviate the omitted variable concerns described in the conceptual framework. On the other hand, these controls may violate the strict exogeneity assumption, since mobility in previous periods may have causal effect on future COVID-19 cases and deaths.

The fixed effect $\alpha_i$ controls for time-invariant characteristics of each county, including typical mobility patterns, rural or urban status, population density, the availability of transportation infrastructure, and all other county-specific drivers of mobility that do not change over time.[8] The date fixed effect $\delta_t$ captures time-varying drivers of mobility and activity, including any national-level event or announcement that influences mobility equality across locations (e.g., announcements from the CDC), pronounced weekday/weekend differences documented in Figure 3.1, and any nationwide weather-driven trend.

Like the previous literature reviewed earlier, the goal of the analysis is to identify the causal effect of MRPs on mobility and activity measures, as represented by the vector of parameters $\beta$. Identification of $\beta$ requires a strict exogeneity assumption for the error term in Equation (1), requiring that it is uncorrelated with the various policy indicators, conditional on the controls and fixed effects included in the regression. This can be interpreted as the standard parallel-trend assumption for the potential outcomes in treated and control counties, whereby the mobility outcomes in "control" counties provide a valid counterfactual for the mobility outcomes in "treatment" counties that adopted MRPs. This assumption can be tested by estimating the event-study analog of Equation (1) and testing for pre-trend differences directly. In the context of COVID-19 policies, the parallel trend assumption could fail due to anticipation of the policies, existence of spillover effects, or reverse causality. Furthermore, when policies are included one at a time, the *interpretation* of $\beta$ as the average treatment effect on the treated relies on assuming that treatment effect is constant across units and/or over time. This paper does not attempt to pin down which of these assumptions

[7] Some counties never experience COVID-19 mortality events in the sample period, in which case the binary indicator remains a zero throughout.

[8] In practice, many of those characteristics may change slowly over time (e.g., population density). For the purpose of this analysis with a sample period that covers a three month period, we consider them time invariant.

actually fail and where, but instead focuses on showing the existence of warning signs in the data that might lead a researcher to reconsider the validity of the identifying assumptions. We return to these important points below.

### 3.4 Results using the TWFE estimator

For the remainder of this paper, we will systematically investigate how researcher's decisions influence the estimated effects of MRPs on mobility. The space of researcher's degrees of flexibility in this context is extremely large. We only focus on a subset of this space, which still includes multiple dimensions: outcome variables, transformations of the outcome variables, combinations of MRPs to consider, included covariates, regression weighting, and estimator used. Due to these many dimensions we cannot present all the results at once. In the following sections, we will present successive "slices" of the multidimensional space of researchers degrees of flexibility, thus highlighting the numerous "forking paths" of the garden as in (Gelman and Loken, 2013).

Table **??** begins the empirical analysis by reporting the TWFE estimates of $\beta_p$ for a log-linear specification with the dependent variable *Completely Home (%)*. The log-linear specification is one of many used in the previous literature, and we also consider alternative specifications further below.

There are 8 columns in Table **??**, each corresponding to a different combination of the MRPs and control variables. In column (1) we consider only state-level MRPs and we find that all three state-level policies (ED, SIPO, and ERC), considered in isolation of the city or county policies, lead to statistically significant increases in the fraction of time spent completely at home. For example, the state SIPO policies are associated with a 4.3% (0.043 log points) increase in the fraction of time spent at home. In column (2) we consider only county-level MRPs and find stronger effects, e.g., a 9.0% increase in the fraction of time spent at home due to a county SIPO. Column (3) considers city-level MRPs. Since the unit of observation for the mobility data is the county, we model the effect of county population share residing in a city with an MRP. The estimated coefficient is positive and statistically significant, indicating that a 10 percentage point increase in a county's population living in a city with mobility reduction policy increases the fraction of time spent at home by 2.3%.

Since the MRPs are often overlapping (e.g., counties with a county SIPO may also have a state ED or SIPO), we then estimate a model where all MRPs are included in the same regression. Column (4) reports the results of this analysis, and shows that estimating the impact of MRPs on mobility separately for state, county, or city-level policies or estimating the effects jointly leads to very similar estimates. Among the state- and county-level binary policies, SIPO policies are the strongest determinants of reduced mobility, as shown

Table 3.3: Two-Way Fixed Effects Estimates of the Impact of MRPs on Log *Completely Home (%)*

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Completely Home (log) | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| State Emergency | 0.008*** | | | 0.010*** | 0.011*** | 0.009*** | 0.013*** | 0.011*** |
| | (0.003) | | | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| State Shelter-in-Place | 0.043*** | | | 0.032*** | 0.033*** | 0.027*** | 0.026*** | 0.022*** |
| | (0.004) | | | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| State Earliest Restriction or Closure | 0.022*** | | | 0.017*** | 0.020*** | 0.016*** | 0.022*** | 0.018*** |
| | (0.005) | | | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) |
| County Emergency | | 0.030*** | | 0.018*** | 0.018*** | 0.010*** | 0.017*** | 0.010*** |
| | | (0.004) | | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| County Shelter-in-Place | | 0.090*** | | 0.066*** | 0.064*** | 0.063*** | 0.055*** | 0.056*** |
| | | (0.008) | | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) |
| County Pop. Share Under City Policy | | | 0.227*** | 0.193*** | 0.188*** | 0.165*** | 0.136*** | 0.122*** |
| | | | (0.014) | (0.013) | (0.013) | (0.013) | (0.013) | (0.013) |
| Temperature (mean) | | | | | −0.001*** | −0.001*** | −0.001*** | −0.001*** |
| | | | | | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Precipitation | | | | | 0.063*** | 0.064*** | 0.063*** | 0.063*** |
| | | | | | (0.001) | (0.001) | (0.001) | (0.001) |
| Snow | | | | | 0.035*** | 0.035*** | 0.035*** | 0.035*** |
| | | | | | (0.001) | (0.001) | (0.001) | (0.001) |
| Cases > 0 | | | | | | 0.084*** | | 0.077*** |
| | | | | | | (0.003) | | (0.003) |
| Deaths > 0 | | | | | | | 0.080*** | 0.069*** |
| | | | | | | | (0.004) | (0.004) |
| Day FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| County FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 285,405 | 285,405 | 285,405 | 285,405 | 282,997 | 282,997 | 282,997 | 282,997 |

*Notes:* Standard errors clustered at the county level. *p<0.1; **p<0.05; ***p<0.01

by the positive coefficients on the log fraction of time spent completely at home. The estimates in column (5) confirm this result and show that controlling for daily weather variables (precipitation, snow and mean temperature) which are also drivers of mobility, does not substantially change the estimated MRPs effects. Columns (6) to (8) complete the analysis by entering indicators for the date at which a county experiences its first COVID-19 case or confirmed death. Again, the addition of these factors does not lead to a meaningful change in the estimated MRP effects.

The evidence in Table **??**, albeit for only one mobility outcome, might lead one to conclude that MRPs had the intended effect by significantly changing mobility behavior through an increase in the fraction of time spent at home. In this case, the estimated MRPs effects are stable and robust across a wide range of specifications, including the joint estimation of MRP effects, and the State SIPO policies are the strongest determinants of reduced mobility. Next, we document that the remarkable stability of the TWFE estimates of the impact of MRPs on log *Completely Home (%)* generally applies to all the other mobility outcomes.

### 3.4.1 Robustness of the TWFE estimates

In this section we continue the analysis of the log-linear specification and graphically report in Figure 3.5 the estimates of the impact of State ED along with their 95% confidence intervals for the 14 mobility and activity indicators that take strictly positive values.[9] Each subplot reports estimates specific to a given mobility outcome. The height of each bar correspond to the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

The subplot in the upper left corner is for the Cuebiq (CQ) *Mobility Index.* We use different colors to represent the specification of the predictors of mobility (detailed in the legend). Starting from the left, the red bar shows the estimate of the effect of State ED on the *Mobility Index* when only the state-level MRPs are included in the model (along with the county and date fixed effects), corresponding to column (1) in Table **??**. The specification shown by the orange bar adds all other policies (county and city-level), and the one in yellow adds the weather controls. The specifications shown with the pale green and darker green add in turn the indicator for the date of the first case or first death related COVID-19 in each county. Finally, the bar to the right (blue) corresponds to column (8) in Table **??** and includes all policies and predictors in the model. For example, in the case of the Safegraph *Median Distance*, we find that the TWFE estimates point to a robust and statistically significant positive impact, indicating that the state ED policies were effective in reducing mobility, by about 0.02 log points, or 2% (recall that variables were re-signed so that positive impacts can be interpreted as increase in social distancing).

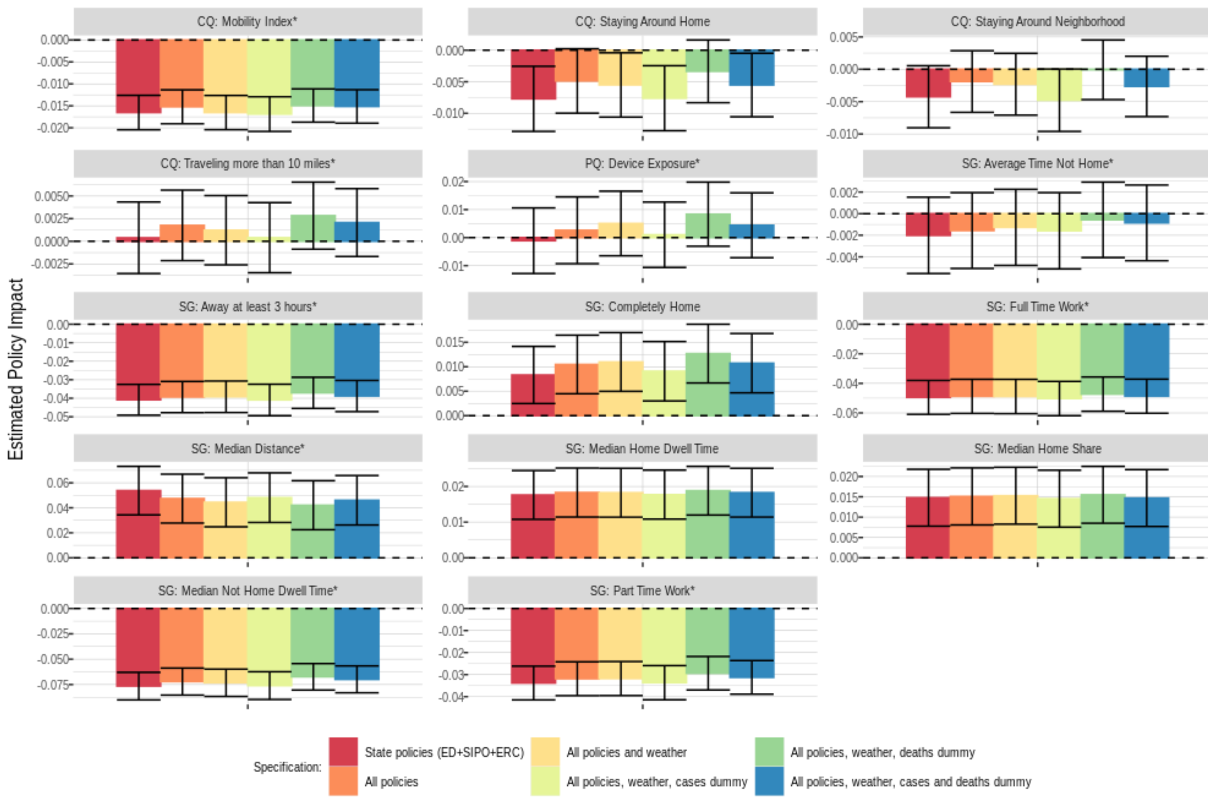
Remarkably, the robustness of the TWFE estimate of the impact of State ED policies on mobility across the log-linear specifications that include in turn all other MRP (state, county, and city) is evident for all 14 mobility outcomes in Figure 3.5. Specifically, for most of the mobility outcomes (i.e., within each individual subplot bar graph), the point estimates are similar and the 95% confidence intervals are generally overlapping, and do not meaningfully change across the 6 specifications.

In the Appendix, we report a series of additional figures for the four remaining MRPs, similarly structured as Figure 3.5 that further document the stability of the TWFE estimates of the impact of MRPs on our 14 mobility outcome modelled with a log-linear specification: State SIPO (Figure F.1), State ERC (F.2), County ED (F.3), and County SIPO (F.4). Similar to the evidence in Figure 3.5, we find that for all MRPs, the TWFE estimates generally have roughly the same rough magnitudes across specifications.

Based on the analysis of a single outcome (for example, log *Completely Home (%)*), it would seem natural to conclude that MRPs played an important role in reducing mobility and increasing social distancing during the first-wave of the COVID-19 pandemic, as other researchers have noted (e.g., Holtz et al. 2020). However, further analysis raises two key concerns. First, the stability of point estimates across specifications in Figure 3.5 is a misleading, or at least incomplete, signal of robustness. While such stability is often used as a basis for establishing a causal relationship, we show later in our event study analysis that assumptions

[9] Since the Google Mobility indicators are reported as a change relative to a reference week and thus contain negative values, we cannot analyze them with a log-linear model. We examine models where the dependent variable is in levels below.

Figure 3.5: Estimated Impact of State ED on Log Mobility Outcomes

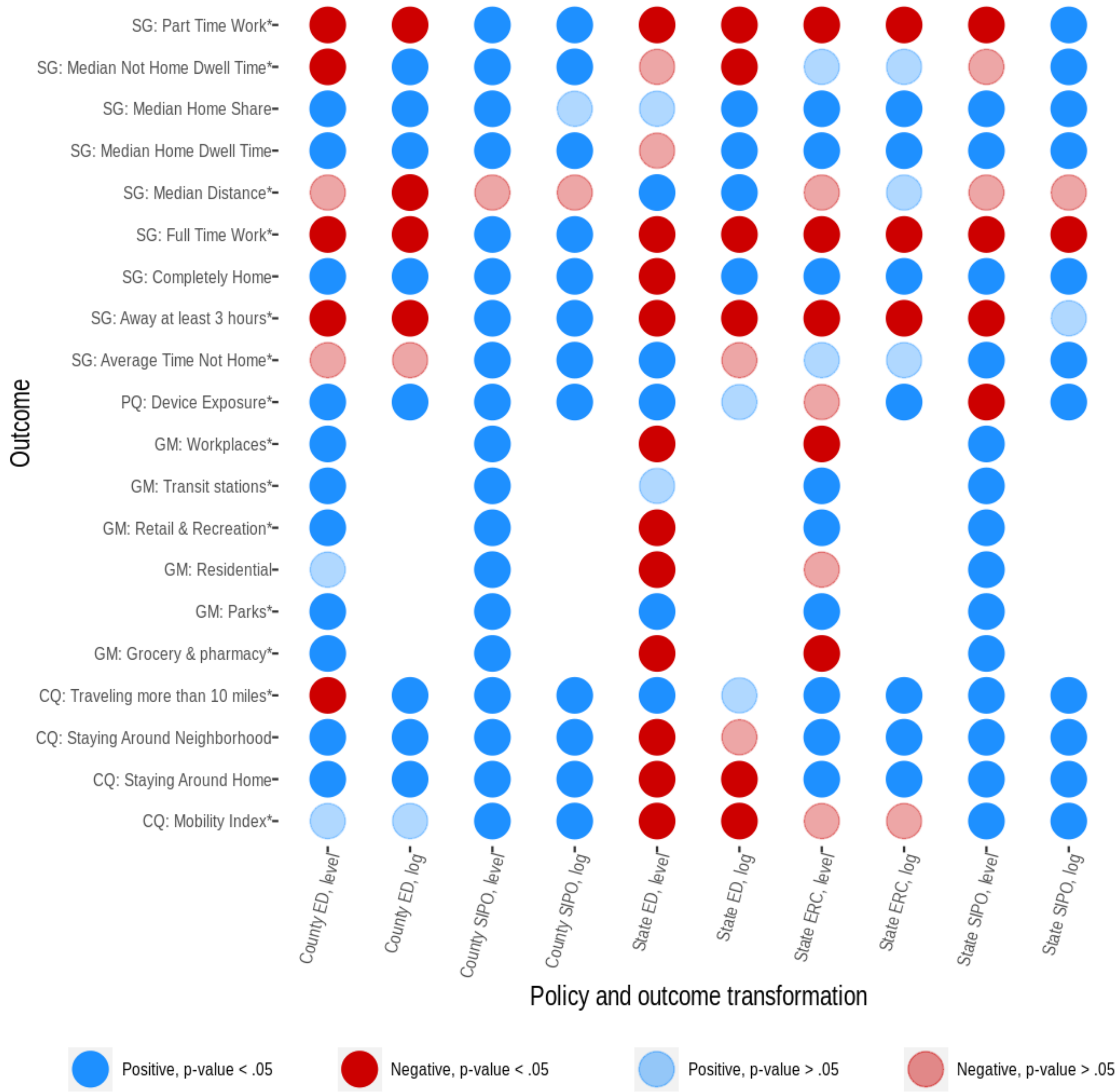



Notes: Figure 7 reports the TWFE estimates of the impact of State ED on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

needed for credible difference-in-difference estimation may not hold. Second, we find persistent issues with the *direction* of our main effect of interest: for several outcomes and policies, the MRPs are estimated to *decrease* distancing rather than increasing it as one would expect. To summarize this evidence visually, Figure 3.6 shows the sign and statistical significance of the estimated coefficient on the MRPs from regressions where the dependent variable enters either in levels, or through a log-transformation (as in Figure 3.5 and all other related figures in the Appendix).

Throughout Figure 3.6, we rely on the "full" specification used in the final column of Table **??** (all policies included simultaneously and including weather and Covid-19 severity controls). We therefore examine multidimensional space of researcher's degrees of flexibility to focus on the impact of the choice of outcome variable and MRP considered, investigating two possible transformations of the outcome. Remember that all dependent variables have been re-signed to indicate mobility reduction, so that we expect MRPs to lead to an increase in social distancing and increase the dependent variables. The rows in the figure correspond to each of our 20 outcome variables while the columns indicate a given MRP and whether the outcome variable

Figure 3.6: Estimated Sign and Statistical Significance of Estimated Impact of MRPs on Mobility Outcomes




Notes: Figure 8 reports the TWFE estimates of the impact of all MRPs on 20 mobility outcomes using either a linear or log-linear specification. All models include all MRPs included simultaneously and the weather and Covid-19 severity controls

is modeled in log or level. Blue dots indicate the expected sign (positive, i.e., more social distancing) while red dots indicate the unexpected sign. For either color, the darker version of the color indicates the estimated MRP coefficient was statistically significant at the 5% level.

Statistically significant coefficients of unexpected sign (shown in dark red) are prevalent in Figure 3.6. This is true for the majority of outcome variables for State ED in both level and log-linear specification. Incorrectly signed impacts are also the majority result across policy and outcome transformation for four outcome variables (*Part Time Work, Median Distance, Full Time Work,* and *Away at Least 3 Hours*). In addition to the prevalence of unexpected signs, we also observe signs switching from statistically significant in one direction to significant in the other direction as we vary either the outcome variable, or the policy and dependent variable transformation. For example, the impact of State SIPO on *Device Exposure* is negative for the specification in levels and positive (as expected) for the log-linear specification, and statistically significant in both cases. Similar sign switches of the estimated impact of the same MRP on the same outcome are observed for *Part Time Work*, *Median Not Home Dwel Time*, *Completely Home*, and *Travelling More than 10 Miles*.

What can explain such pervasive reversal of estimated MRPs impact across specifications? A first possibility comes from the recent econometric literature documenting that when treatment adoption is staggered and when there is heterogeneity between cohorts of units that become treated at different times, the TWFE estimator fails to recover a meaningful average treatment effect (de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; Sun and Abraham, 2020; Borusyak et al., 2021). As shown in Figure 3.3 all MRPs analyzed in this paper and in the previous literature were implemented in a staggered manner. Furthermore, counties that implemented MRPs at different times were likely impacted differently by these policies, which could introduce between adoption cohorts treatment effect heterogeneity. Thus a first consideration is that the estimates reported in the previous section might fail to recover the ATT due to issues inherent to how the TWFE estimator handles treatment effect heterogeneity.[10]

A second possibility is that the parallel trend assumption—one of the central assumptions underlying the identification of ATTs in the difference-in-differences framework—is invalid for the potential outcomes of some variables. In particular, in the case we highlighted earlier where the impact of state ED policies on the *Completely Home (%)* variable in the log-linear specification is positive and statistically significant but negative and statistically significant in the specification in levels could indicate that the parallel trend assumption holds in the log-linear specification, but not in the level specification (or vice versa).[11] Recent work by Roth and Sant'Anna (2021) highlight that in general, the parallel trend assumption is sensitive to the specific functional form chosen in the estimating equation. In the context of the mobility impacts of COVID-19 policies, economic theory does not provide much guidance to researchers regarding the choice of

[10]In subsection I in the Appendix, we perform the Goodman-Bacon decomposition that highlights how the implicit use of counties treated earlier as controls for units treated later can be problematic. In particular, we find that removing all comparisons that involve the use of counties treated earlier as control units from the TWFE estimator frequently changes the *sign* of the estimated effect for the State ED and State ERC policies (the two MRPs that do not have a well-defined control group throughout our sample period). However, the sign and magnitude of the estimated effect remain stable for most other outcome variables, suggesting that a different assumption might also be violated.

[11]This concern over possible failure of the parallel trend assumption is heightened by the observation that many individuals may have curtailed their mobility behaviors even if their state or county of residence was not under an MRP, or in anticipation to the pre-announcement of a SIPO or other MRPs. Rambachan and Roth (2021) provide a way to relax the parallel trend assumption by restricting the magnitude of post-treatment violations of parallel trends.

functional form. Unless a specific transformation can be theoretically motivated, using various functional forms for the outcome variable can provide one way to assess robustness. The next section presents a detailed analysis of pre-trends for all the mobility outcomes and MRPs.

## 3.5 Event study analysis

### 3.5.1 Specification

The DD analysis of the impact of MRPs on the mobility outcomes presented in the previous section relies on the parallel trend assumption, which states that trends in the outcome (mobility indicator) before the adoption of an MRP (known as "pre-trends") are the same in the treatment (adopters) and control (non-adopters) counties. An event study analysis provides a simple graphical approach at documenting pre-trends, which we now implement and supplement with F-tests on the pre-trend coefficients in order to test the parallel trend assumption.

To proceed, we use the same framework as in Equation (3.1), but index days relative to the event (date of adoption of an MRP) by $k$. Therefore $D_{ipk}$ is a dummy variable equal to 1 when county $i$ is $k$ days away from being "treated" by a given MRP policy $p$ within the set $P$. As before, the policy set is $P = \{$state ED, state SIPO, state ERC, county ED, county SIPO$\}$. Formally, $D_{ipk} = 1\{t - MRP_{ip} = k\}$, where $MRP_{ip}$ is the policy $p$ for county $i$. We take the common event study approach of including a single dummy for all relative days before our event window (over 20 days pre-adoption of a policy) denoted by $k = -21^-$, and another for all relative days after, $k = 21^+$ (over 20 days post-adoption):

$$Y_{it} = \alpha_i + \delta_t + \left( \sum_{p \in P} \sum_{k=-21^-}^{k=21^+} \theta_{pk} \cdot D_{ipk} \right) + X'_{it}\gamma + \epsilon_{it} \tag{3.2}$$

The day before the adoption of a policy ($k = -1$) is omitted to serve as baseline. Like in the standard DD analysis, we include county fixed effects ($\alpha_i$) to control for unobserved time-invariant differences between counties, and day fixed effects ($\delta_t$) to control for unobserved time-varying drivers of mobility and activity that are common to all counties. We also include the full set of time-varying control variables ($X_{it}$) we considered earlier (see last column in Table **??**.) Finally, $\epsilon_{it}$ is a county-day specific error term. Standard errors are clustered at the county level.

### 3.5.2 Results from the event study analysis

To begin, we focus on a single outcome, *Completely Home (%)* and graphically present estimates of the event-study $\theta_{pk}$ coefficients (along with the corresponding 95% confidence intervals) in Figure 3.7. We consider 8 different possible specifications of the dependent variable, including the level and log-transformed specifications analyzed earlier. We also consider specifications where the dependent variable is first-differenced, and specifications where the dependent variable is a 7-day moving average of the mobility outcome (as opposed to the daily-level value). We either weight the underlying regressions by county-level population, or estimate them without weights. We implement these 8 specifications because they have been used in the previous literature. The labels on the right-end margins of each row in Figure 3.7 indicate the specification of the dependent variable and use of weights. The 5 columns in Figure 3.7 correspond to the five MRPs we analyze. Note that in each column, we report the estimated event-study coefficients for the relevant MRP estimated in isolation (color) and in combination with the four other MRPs (black).[12]

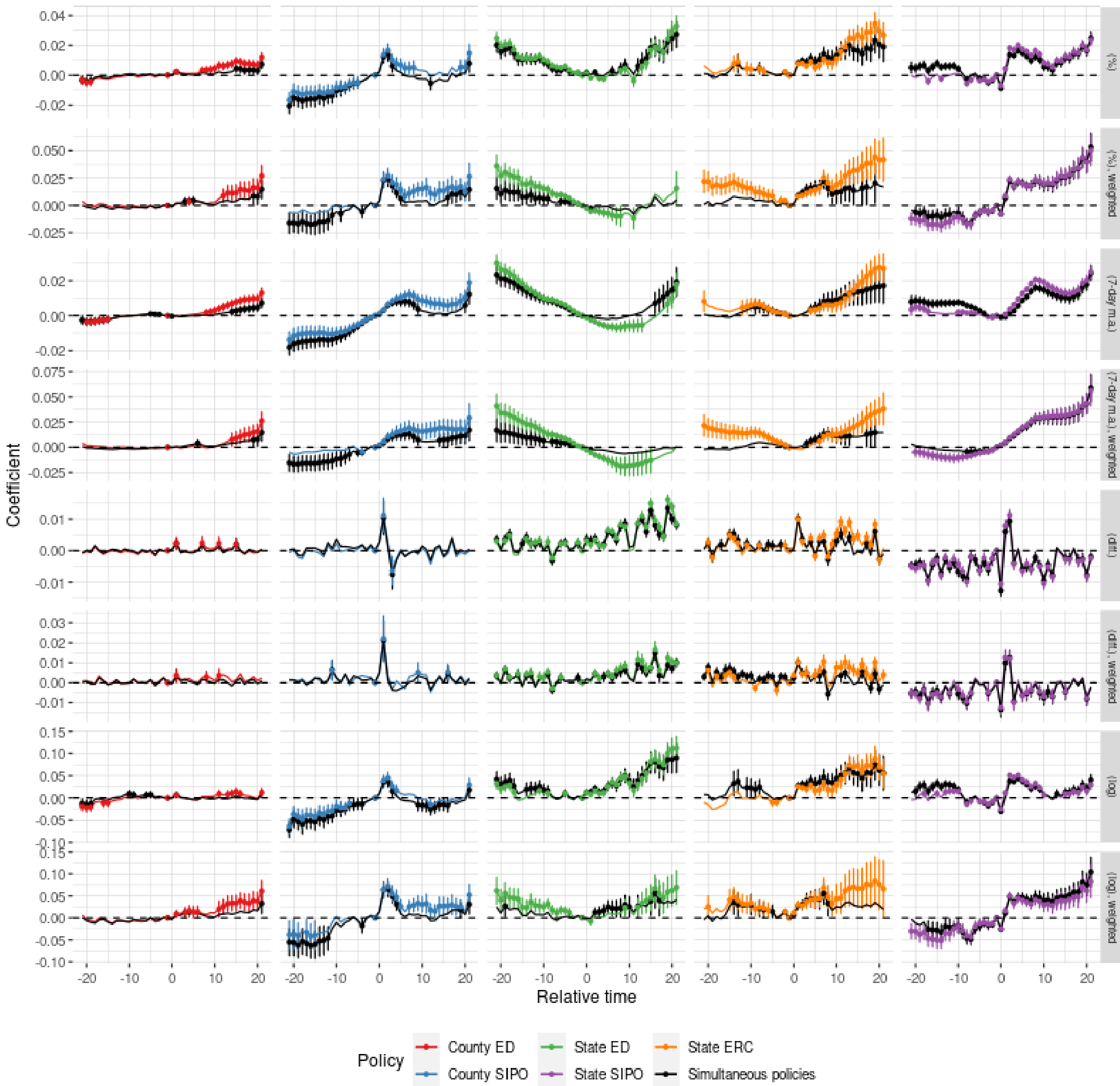
There are three main results that readily emerge from Figure 3.7. First, focusing on the post MRP adoption period (to the right of 0 on the x-axis), most specifications and MRPs show results that are consistent with MRPs increasing the *Completely Home (%)* variable (although there are a few exceptions such as the 7-day moving average model for state ED and the first-differenced models for state SIPOs). Second, with the exception of county ED, statistically significant and sizable pre-trends are apparent for most specifications and MRPs. Third, including all policy variables simultaneously (black estimates) as opposed to individually, using population weights, or first-differencing the outcome does not lead in general to reductions in the magnitude of the pre-treatment coefficients.

To assess whether the issues with pre-trends and instability of the estimates identified previously could simply be due to our choice of dependent variable, we present additional slices of the degrees of flexibility space below. Figures 3.8 and Appendix Figure F.10 continue the event-study analysis by reporting estimates for all mobility outcomes, for the two initial specifications of the dependent variable (log-linear and in level). Each box in the figures correspond to a given mobility outcome and shows the estimated event-study coefficients ($\theta_{pk}$) and 95% confidence intervals for the 5 MRPs, each from a separate regression.

Examining first the results from the specification in levels for the outcomes, one can observe the complex dynamics in the impact of MRPs on mobility and activity outcomes in the post adoption period, as was also shown in Figure 3.7. Some policies have rapid impacts in decreasing mobility, with impacts lasting

[12] To the best of our knowledge, only a couple of papers have studied the identification of treatment effects under multiple treatment. In an application focused on the impacts of school bonds on housing prices, Cellini et al. (2010) derive estimators of the "treatment-on-the-treated" and "intent-to-treat" when multiple bonds measures have been passed, assuming that dynamic effects only depend on the length of time elapsed since treatment. Using simulation studies, Sandler and Sandler (2014) highlight that in the case of multiple treatment, ignoring one or more treatment produces biased estimates in general. Two working papers recently released highlight additional issues when estimating the impacts of multiple treatment (Goldsmith-Pinkham et al., 2021; de Chaisemartin and D'Haultfœuille, 2022).

Figure 3.7: Estimated Event-Study Coefficients for *Completely Home (%)*




Notes: Figure 9 reports the event-study estimates of $\theta_{pk}$ from equation (2) above for the *Completely Home (%)* outcome under several specifications. The transformation and use of regression weights varies by row and the MRP varies by column. The estimates are presented for both the MRP estimated in isolation (shown by the colored lines) and jointly with all MRPs (shown by the black lines). Whiskers represent the 95% confidence intervals.

from a few days, to the full 20 days in the post-adoption period. Notably, some of the post MRP adoption estimates have the "wrong" sign, indicating that the introduction of an MRP increased mobility and reduced social distancing (e.g., the state ED effect on Safegraph *Away at least 3 hours*, or the county SIPO effect on Safegraph *Median Home Dwell Time* and *Median Home Share (%)*). At the same time, it is also surprising to observe that the different MRPs can have differently signed impacts on the same mobility outcomes (e.g., Safegraph *Full Time Work*), which is increased by State SIPO and reduced by State ED.

Most importantly, however, it is again evident that sizable and statistically significant pre-trend differences exist for at least one policy for each outcome. This adds to the evidence that *at least one* of the assumptions required to interpret event-study coefficients as estimates of the ATTs is violated (that is, homogeneity of the treatment effect, parallel trends for the potential outcomes, and no anticipation). The nearly linear trend in pre-treatment coefficient also suggests the "under-identification" issues that arise due to the no-anticipation assumption not being explicitly imposed, as highlighted in Borusyak et al. (2021). Looking at all available mobility outcomes and for most policies analyzed, a large number of pre-trend coefficients are statistically different from zero at the 5% level. For example, for the Cuebiq *Travelling More Than 10 Miles* outcome, we can easily detect significant pre-trends for all five MRPs. Moreover, the pre-trend patterns tend to vary across policies, again in the case of Cuebiq *Travelling More Than 10 Miles* outcome, the pre-trend coefficients are mostly positive for state ERC and county ED, while negative for state ED, State SIPO, and County SIPO.

Figure F.10 in the Appendix is structured similarly as Figure 3.8 but focuses on models with a specification in levels, which includes mobility outcomes from Google [13] For all outcomes, we detect statistically significant pre-trends for one or more of the policies. Moreover, for any given outcome, the pre-trends can be consistently positive or negative, depending on the MRP considered. This heterogeneity in the impact of different MRPs on the same outcome is also observed in the post-adoption period.

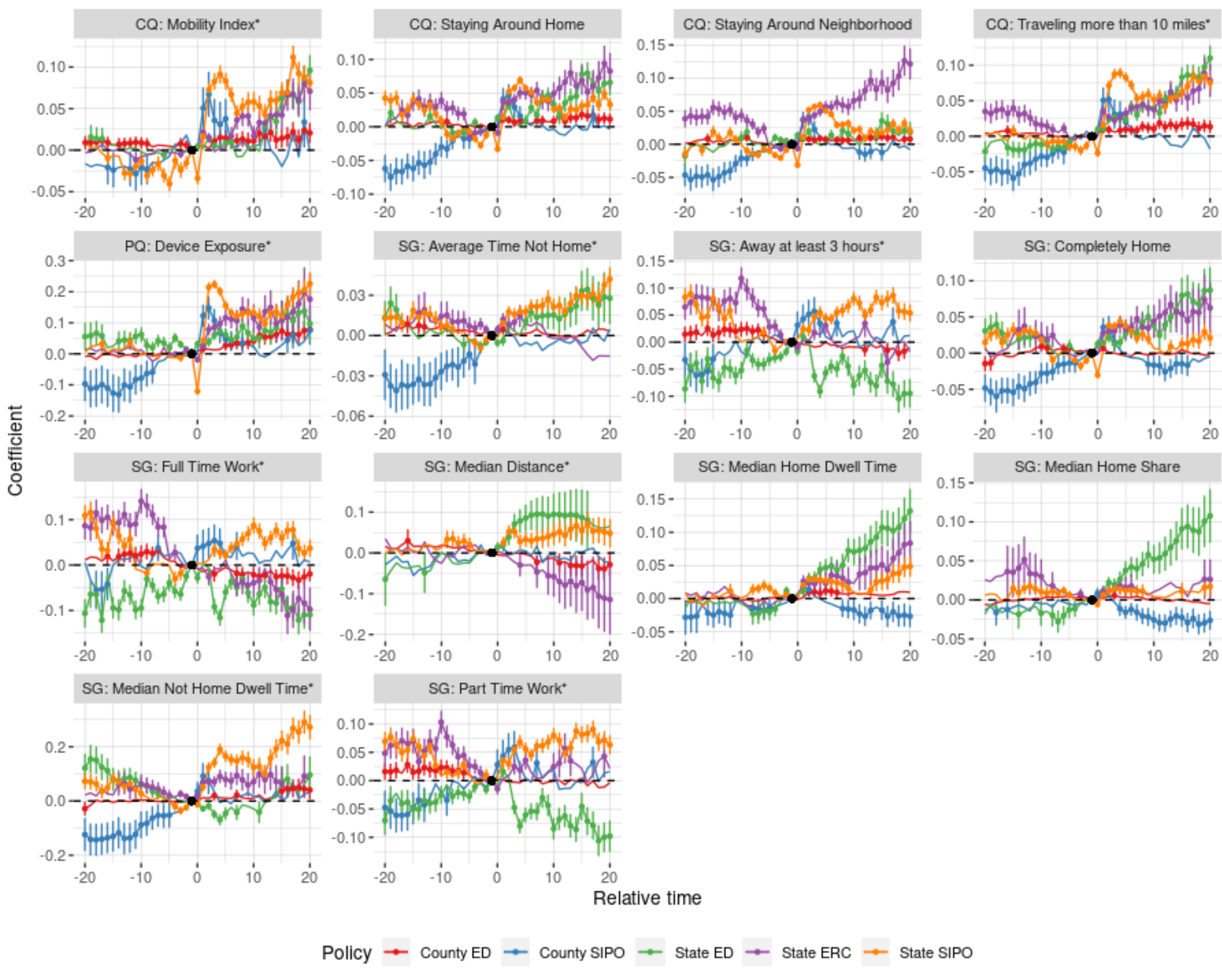
Figure F.11 summarizes the information about the pre-trends across the 8 specifications considered in Figure 3.7 and all outcomes. The figure reports the p-values from F-tests testing the null hypothesis that the pre-MRP-adoption coefficients are jointly equal to zero. It is configured as a heat-map, with the columns indicating the five MRPs estimated individually and then estimated jointly (far right column). The rows correspond to the eight specifications for the dependent variable that were considered in Figure 3.7. Each cell summarizes the results of 14 F-tests on pre-trends coefficients, corresponding to the 14 mobility outcomes we analyze from Safegraph, Cuebiq, and PlaceIQ.[14] The number in each cell corresponds to the fraction of tests (out of 14) where the p-value on the null hypothesis exceeds 0.05, indicating that the null hypothesis would be rejected at the usual 5% significance level.

The results further confirm some of the evidence already documented: significant pre-trends are pervasive. For example, *all* 14 out of 14 "no pre-trends" tests are rejected for all specifications of the dependent variable for the state ED, state SIPO, state ERC, and simultaneous policies models (as shown with a fraction of zero in the cells). For county SIPO, the parallel trend assumption is rejected for all 14 outcomes in half of the specifications (first-difference, first-difference weighted, log-linear weighted, and 7-day moving average

[13] Recall that due to the log-linear specification of the model, outcomes from Google Mobility are not included since they can on zero and negative values.

[14] The Google Mobility outcomes are ignored here since we cannot apply the log transformation on them.

Figure 3.8: Estimated Event-Study Coefficients for All Mobility Outcomes, Based on the Log-Linear Specification




Notes: Figure 10 reports the event-study estimates of $\theta_{pk}$ from equation (2) above for 14 mobility outcomes, using the log-linear specification. Each estimated impact of MRPs are estimated from separate regressions. Whiskers represent the 95% confidence intervals.

weighted). The MRP for which the parallel trend assumption appears most supported is county ED, although we fail to rejected at most for 6 of the 14 outcomes in 1 specification. All together, 672 F-tests on pre-trends coefficients were conducted to construct Figure F.11. Out of those, we fail to reject the null hypothesis of "no pre-trends" only 34 times (out 672 tests) at the 5% significance level.

Differences-in-differences and event-study approaches, implemented with standard TWFE estimators to the mobility outcomes to estimate the impact of MRPs provides definitive evidence of estimates that are "robust" to the inclusion of covariates. But we also observe consistent violation of the parallel trends assumption—rejection of the null hypothesis of no pre-trends—and major specification issues where the sign of the effect depends on choices made on the transformation of the dependent variable, e.g., linear versus log-linear versus first-difference. This combined pattern of stability of estimates within-specification, but instability across specifications (for the same outcome), and statistically significant pre-trends is observed for most outcomes considered. We interpret this evidence that one or more of the assumptions required to interpret the DD and event-study estimates as Average Treatment Effects on the Treated (ATT) are likely violated.

## 3.6 Heterogeneity-robust estimators for staggered adoption designs

The recent econometric literature highlights that under the presence of staggered treatment and treatment effect heterogeneity, the TWFE regressions underlying DD and event-study analyses typically fail to recover the ATT (Sun and Abraham, 2020; de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021). The bias arises from the implicit construction of control groups which include units that are themselves under the effect of a treatment.[15] For example, in our setting, TWFE regressions will *implicitly* use counties treated on March 15 to serve as controls for counties treated on March 20; this might be a bad comparison if the treatment effect is dynamic, or if the treatment effect is heterogeneous across units treated on different days. This same literature also provides various solutions to address the deficiencies of the TWFE estimators. New event-study methods, for example from Callaway and Sant'Anna (2020), reduce the problem to estimating "clean" difference-in-differences for each group of units that receive treatment at the same time (also called "cohorts") by comparing these units to proper "controls" (which are units either not yet treated, or never treated). The cohort-specific difference-in-differences estimates can then be re-aggregated to produce event study estimates that are free of the bias due to using treated units as implicit controls.[16]

Applying the Callaway and Sant'Anna (CS) estimator in our setting with the control variables leads to several observations being dropped, sometimes to the extent that a given model will not be estimated.[17] With this as a limitation, we only consider CS estimates that consider one MRP at a time. In addition, because the CS estimator and related estimators only compare units that are treated to units that are never or not-yet-treated, this explicitly restricts the number of post-treatment periods for which we can estimate

[15] Goodman-Bacon (2021) shows this issue with a decomposition result that reframes the TWFE estimator as a weighted average of all possible two-by-two difference-in-differences estimators. Sun and Abraham (2020) propose a general decomposition of the event study estimator for alternative choices of specifications that nests TWFE as a special case.

[16] de Chaisemartin and D'Haultfœuille (2020a) propose a similar estimator motivated from the point of view of a social planner who could discount future treatment effects, à la Manski (2005). Sun and Abraham (2020) propose an alternative estimator obtained by interacting event-study coefficients with dummies for each treatment cohort. Both alternatives weight cohort-specific treatment effects by relative cohort size.

[17] This occurs because the CS estimator requires a certain overlap of the categorical covariates between treatment and control groups. While the overlap holds over the whole sample, it can fail to hold in some subsamples of smaller size. In that case, the problematic subsample is simply dropped.

a treatment effect. In our sample all units are eventually treated by a State ED and a State ERC, and the last unit to receive a state ED does so on March 16. This means that no treatment effect of State ED can be estimated after March 16 with the CS estimator. To ease comparison with the previous results in the paper, we report CS estimates focusing on the three MRPs for which there are always proper "control" units during the sample period, namely County ED, County SIPO, and State SIPO.

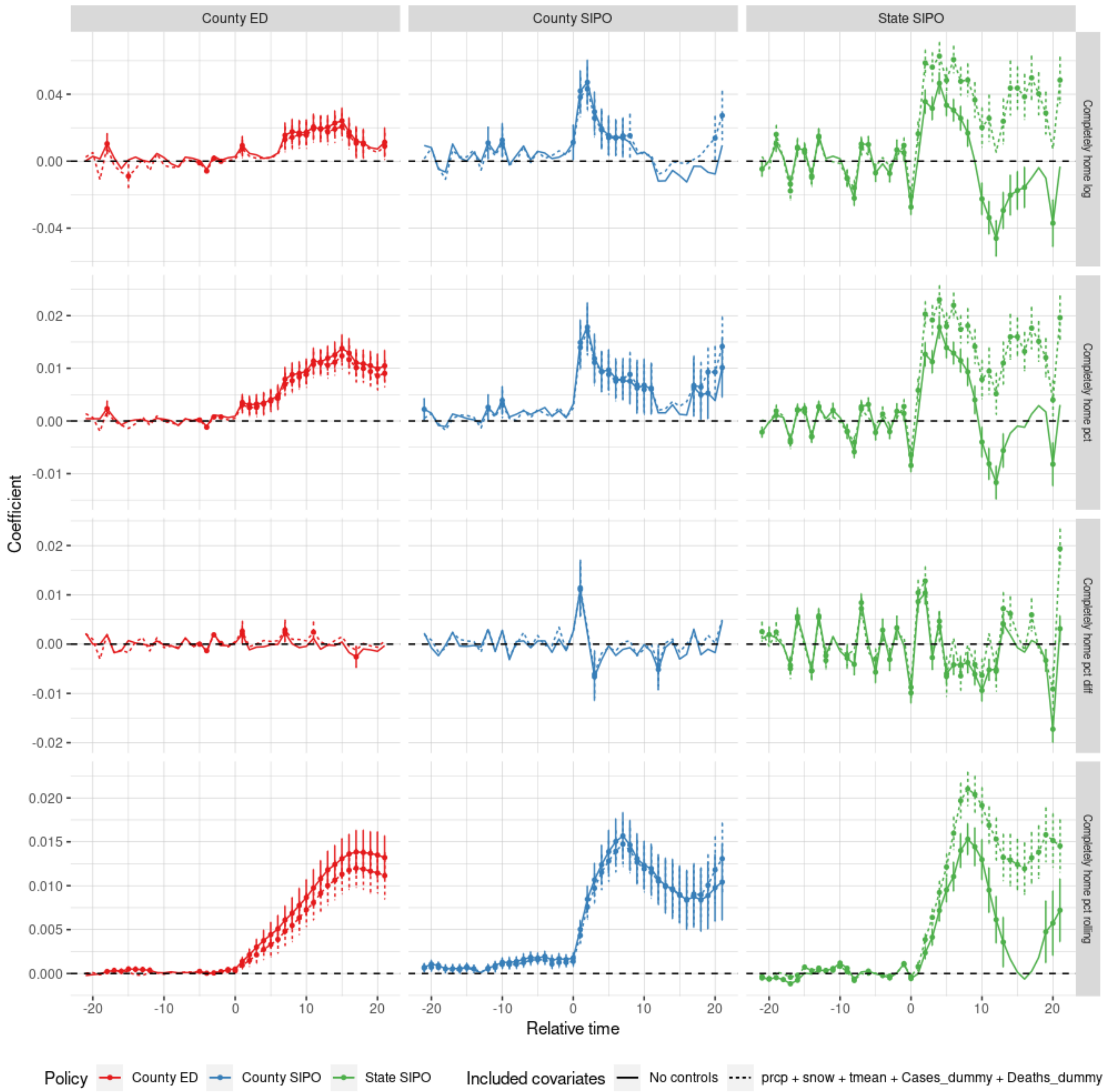
Figure F.12 in the appendix reports the CS estimates for the cohort-specific impact of these three MRPs on the *Completely Home (%)* mobility outcome. These CS estimates are analog to the standard event-study regression specification results presented above in Figure 3.7. Here, each MRP cohort (defined as each individual MRP adoption date in the sample) is represented by a separate line and color. We continue to present estimates for various transformation of the dependent variable as the previous sections have shown how seemingly minor specification changes (e.g., linear versus log-linear) can lead to widely different estimates. The estimates in Figure F.12 are noisy, as might be expected since they represent around 50 smaller, cohort sub-samples from the overall data set.

Figure 3.9 is structured as Figure F.12 but combines the cohort-specific estimates into a single estimate using the aggregation method proposed in Callaway and Sant'Anna (2020) to obtain heterogeneity-robust event-study estimates of the ATTs. Again, we focus on the *Completely Home (%)* mobility outcome for illustrative purposes. For each MRP and transformation of the dependent variable, two set of estimates are reported: with and without covariates. While the aggregated CS estimates still point to important pre-trend differences, especially for County and State SIPO, the overall picture is better than in the case of standard event study estimates. Notably, the pre-event coefficients are small relative to post-treatment coefficients.

Taken together, the CS estimates in Figure 3.9 for *Completely Home (%)* be a lot more stable than TWFE estimates, but still offer mixed results regarding the impact of MRPs on time spent completely at home. In the case of County ED, the estimates point to a positive impact on time spent completely at home (except for the specification using first differences of the outcome in row 3). Similar patterns are observed for County SIPO, although the estimates are noisier and pre-trend differences are more noticeable. For State SIPO the estimates are more dependent on the inclusion or exclusion of the covariates (weather variables and indicators for county-level COVID-19 cases and mortality).

Among all specifications considered, the CS event study estimates for the 7-day moving average transformation appear most consistent with the "no pre-trend" hypothesis. As shown in Figure F.13 in the Appendix, this holds across most outcome variables, although unexpected signs still appear for a subset of outcome variables, especially looking at the impacts of County ED. This suggests that—in combination with the heterogeneity-robust estimation—temporal smoothing over a longer periods longer than 1 day helps the credibility of the estimator. In contrast, regressions using the CS estimator still appear very noisy when the

Figure 3.9: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts




outcome is in level, in log, or first-differenced (see Figures F.15, F.14, and F.16 in the Appendix respectively). One implication is that time-aggregating the data may help improve the credibility of the parallel-trend assumption. Notably, a similar finding emerges from the previous literature where event studies conducted with state-level data (as opposed to county-level) generally produce well-behaved estimates (e.g., Andersen 2020; Abouk and Heydari 2021; Dave et al. 2021). Thus, it may be that fine scale event study estimates in our setting (daily at the county level) are beyond the capacity of the data and difference-in-difference based methods, unless one is willing to relax the parallel-trend assumption.

## 3.7 Conclusion

In this article we analyze the impact of mobility reducing policies (MRPs) on mobility outcomes during the onset of the COVID-19 pandemic, leveraging a uniquely broad coverage of policies spanning the state, county, and city levels in the United States. Generalizing from previous research which focused on selected mobility outcomes and used specific functional forms and TWFE estimators, we analyze twenty mobility measures derived from aggregated mobile device signal data using a wide range of specifications and estimators. We also analyze five different types of MRPs both in isolation and jointly, consistent with the manner they were implemented at the state, county, or city level.

In this context, we find that the standard TWFE and event-study estimates are highly sensitive to choices of mobility outcome and transformation of the outcome (e.g., linear versus log-linear) being implemented. In particular, we uncover systematic violations of the standard parallel-trend assumption for virtually all outcomes and MRPs. We also document that due to researcher's degrees of flexibility, it is possible to focus on specific outcome variables and functional forms that produce "well-behaved" estimates, and that usual robustness-tests based on the sequential addition of covariates points to the stability of those estimates. Yet, for several outcomes we obtain widely different results for the impact of MRPs on the same outcome (often with differing sign), but analyzed through TWFE estimator on different functional forms. Due to the absence of guidance from economic theory to motivate the choice of a specific functional form in this context, one could conclude that both the magnitude and the sign of the ATT associated with the effect of mobility reducing policies on mobility remains largely unknown.

An emerging econometric literature highlights that in the presence of treatment effect heterogeneity and staggered treatment adoption, standard difference-in-differences and event-study estimators can be severely biased (Goodman-Bacon, 2021; Sun and Abraham, 2020; de Chaisemartin and D'Haultfœuille, 2020b; Borusyak et al., 2021). Indeed, we hypothesize that the unreliability of the basic TWFE estimates we uncover is likely due to heterogeneity in the treatment effect of MRPs between groups of counties that receive treatment at different times. When we implement the recent estimator by Callaway and Sant'Anna (2020), we generally obtain more stable estimates of the overall ATT of MRPs, where the direction of the estimated impact is more robust across specifications. At the same time, these important new methods have noteworthy limitations. In particular, we find that due to the interaction and overlap between different MRPs, heterogeneity-robust difference-in-differences estimators fail to credibly estimate the impacts of multiple treatments. Further, several of the estimated pre-MRP coefficients (i.e., "pre-trend" coefficients) are statistically different from zero using these more recent estimators, which weakens the case for a causal interpretation of the estimates. Finally, the transformation of the outcome variable still has large impact on the estimated ATT, sometimes changing the sign of the estimates.

Taken together, there are three main implications to these results. First, since the onset of the pandemic, dozens of research articles used standard difference-in-differences and event-study methods and broadly concluded that MRPs caused reductions in mobility, suggesting they were an important tool to curb the spread of COVID-19. In addition, a related literature used similar methods to estimate the impact of MRPs on COVID-19 cases and deaths. Our reading of the evidence suggests the previous literature appears to have focused on a set of specifications that are highly sensitive to varying minor modelling decisions. Indeed, most prior studies of the impact of MRPs found that such interventions reduce mobility, but this conclusion requires the researcher to favor one specification over other plausible ones. As we find here, for most mobility outcomes, there exist alternative specifications that provide contrary results. Overall, this suggests that robustness of recent evidence of a causal impact of MRPs on reducing mobility is not unequivocal.

Second, a lot of recent attention in the econometric literature concerned with staggered treatment adoption designs has focused on deriving treatment effect heterogeneity-robust estimators. Our empirical results indicate that such estimators can produce estimates that are substantially different than those relying on standard DD methods (this point is also highlighted in Baker et al., 2021), and indeed improve the credibility of the ATT estimates. However, our research highlights important limitations of these new estimators. To begin, they currently do not account for multiple, potentially interacting treatments. This is a serious limitation in the context of COVID-19, where states, counties, and cities all implemented different kinds of MRPs. More generally, this limitation applies in other settings, such as the study of federal and state overlapping policies (e.g., air quality regulations, education policy, health insurance programs for low-income families). Furthermore, causal identification with these estimators continues to rely on a parallel-trend assumption, and the validity of such assumption usually depends on a specific functional form specification for the dependent variable (Roth and Sant'Anna, 2021). When the chosen functional form substantially impacts the treatment effect estimates or the existence of a “pre-trend”, there is little guidance from theory to select a specific transformation. The Change-in-Changes model developed by Athey and Imbens (2006) provides an alternative approach that does not depend on the scale of the outcome variable, but this estimator currently does not easily accommodate the inclusion of covariates and of many different treatment groups. Continuing to improve and generalize methods for staggered treatment adoption designs and multiple treatments is therefore an important avenue for future research.

Third, although difference-in-differences research designs are straightforward to implement with readily available data and can increase the credibility of observational studies (Angrist and Pischke, 2010), our findings highlights that researchers' degrees of flexibility can still have a marked influence on the sign and magnitude of the estimated treatment effects. We explore the role of researchers' degrees of flexibility along four different margins: choice of outcome variable, functional forms for the dependent variable, covariates, and estimation procedures. This is only a subset of the margins along which degrees of flexibility exist.[18] Estimating models for the full-set of possible specifications is often impossible, and researchers' degrees of flexibility in themselves are not to be abhorred, as they are an irreducible part of research. However, the scale of available discretion in choices made by researchers and their impacts on the estimates, even after conditioning on a "credible research design", raises the question of when to believe the reported estimates. The problem of pre-testing specifications and reporting only a set of results that "fits" within the researcher's overall narrative is well-known (Christensen and Miguel, 2018; Kasy, 2021). To mitigate it, Leamer (1983) argued that applied economists should report the full set of specifications that they estimated, not just the ones that "worked". Four decades later, his recommendation continues to be largely ignored. Even-though robustness tests are pervasive in the applied economics literature, we find that it is possible that robustness holds along one margin, while (consciously or unconsciously) ignoring that robustness fails along another dimension.

In light of our findings, instead of reporting robustness tests that work, one possibility could be to report robustness that are increasingly strict until they fail. The relevant question is not whether the results are robust, but instead "*how much* can we depart from the preferred choices before the main findings no longer hold?". In the context of COVID-19 research, this breaking-point threshold is very low, especially when using standard TWFE regressions. Due to the urgency of obtaining actionable policy results, these shortcomings were not fully appreciated by previous research, including our own. The econometric work highlighting issues with common specifications developed extremely fast and in parallel with the applied COVID-19 research. Therefore, further developments are still needed to make rigorous causal inference about the effects of COVID-19 policies.

[18]For example, additional margins of flexibility include the number of lags and leads in event studies, the definition of the treatment variables, or the choice of the temporal window used in the analysis.

## 3.8 References


Abouk, R. and B. Heydari (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the united states. *Public Health Reports 136*(2), 245–252. PMID: 33400622.

Aiken, A. M., C. Davey, J. R. Hargreaves, and R. J. Hayes (2015, 07). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *International Journal of Epidemiology 44*(5), 1572–1580.

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Yang (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics 191*, 104254.

Andersen, M. (2020). Early evidence on social distancing in response to COVID-19 in the united states. *Available at SSRN 3569368*.

Angrist, J. D. and J. S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Athey, S. and G. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–497.

Baker, A., D. F. Larcker, and C. C. Y. Wang (2021). How much should we trust staggered difference-in-differences estimates? *Working Paper*.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *Working paper*, 48.

Botvinik-Nezer and many authors (2020, Jun). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature 582*(7810), 84–88.

Callaway, B. and T. Li (2021, May). Policy evaluation during a pandemic. *ArXiv*, 1–37.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics In Press.*

Cellini, S., F. Ferreira, and J. Rothstein (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics 1*, 215–261.

Chernozhukov, V., H. Kasahara, and P. Schrimpf (2021). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the u.s. *Journal of Econometrics 220*(1), 23 – 62. Themed Issue: Pandemic Econometrics / Covid Pandemics.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

Clemens, M. A. and J. Hunt (2019). The labor market effects of refugee waves: Reconciling conflicting results. *ILR Review 72*(4), 818–857.

Couture, V., J. Dingel, A. Green, J. Handbury, and K. Williams (2020). Exposure indices derived from placeiq movement data. https://github.com/COVIDExposureIndices/COVIDExposureIndices. Accessed: 2020-05-01.

Dave, D., A. I. Friedson, K. Matsuzawa, and J. J. Sabia (2021). When do shelter-in-place orders fight COVID-19 best? policy heterogeneity across states and adoption time. *Economic Inquiry 59*(1), 29–52.

de Chaisemartin, C. and X. D'Haultfœuille (2020a). Difference-in-differences estimators of intertemporal treatment effects. *Working Paper*.

de Chaisemartin, C. and X. D'Haultfœuille (2020b, September). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

de Chaisemartin, C. and X. D'Haultfœuille (2022). Two-way fixed effects regressions with several treatments. *ArXiv*, 1–34.

Elenev, V., L. Quintero, A. Rebucci, and E. Simeonova (2021, July). Direct and spillover effects from staggered adoption of health policies: evidence from covid-19 stay-at-home orders. *NBER*, 48.

Foote, C. L. and C. F. Goetz (2008). The impact of legalized abortion on crime: Comment. *The Quarterly Journal of Economics 123*(1), 407–423.

Fullman, N., B. Bang-Jensen, G. Reinke, B. Magistro, R. Castellano, M. Erickson, K. Amano, J. Wilkerson, and C. Adolph (2020). State-level social distancing policies in response to COVID-19 in the US. http://www.covid19statepolicy.org. Version 1.98, November 9, 2020.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem,even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2021). On estimating multiple treatment effects with regression. *ArXiv*, 1–26.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics In press.*

Goodman-Bacon, A. and J. Marcus (2020). Using difference-in-differences to identify causal effects of covid-19 policies. *Survey Research Methods 14*(2), 153–158.

Goolsbee, A. and C. Syverson (2021). Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of Public Economics 193*(104311).

Gupta, S., T. D. Nguyen, F. L. Rojas, S. Raman, B. Lee, A. Bento, K. I. Simon, and C. Wing (2020). Tracking public and private responses to the COVID-19 epidemic: evidence from state and local government actions. *National Bureau of Economic Research.*

Holtz, D., M. Zhao, S. G. Benzell, C. Y. Cao, M. A. Rahimian, J. Yang, J. Allen, A. Collis, A. Moehring, T. Sowrirajan, et al. (2020). Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences 117*(33), 19837–19843.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, T. Pugatch, M. Saavedra, and Y. Stopnitzky (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry n/a*(n/a), 1–17.

Kasy, M. (2021). Of forking paths and tied hands: selective publication of findings, and what economists shoudl do about it. *Journal of Economic Perspectives 35*(3), 175–192.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*, Volume 53. Wiley New York.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*(1), 31–43.

Manski, C. F. (2005). *Social choice with partial knowledge of treatment response.* Princeton University Press.

National Association of Counties (NACo) (2020). COVID-19 pandemic response: County declaration and policies. https://www.naco.org/resources/featured/counties-and-covid-19-safer-home-orders. Accessed: 2020-04-15.

Painter, M. and T. Qiu (2021). Political beliefs affect compliance with government mandates. *Journal of Economic Behavior & Organization 185*, 688–701.

Rambachan, A. and J. Roth (2021, May). An honest approach to parallel trends. *Working paper*, 1–126.

Roth, J. (2019). Pre-test with caution: event-study estimates after testing for parallel trends. *Working Paper*.

Roth, J. and P. H. C. Sant'Anna (2021). When Is Parallel Trends Sensitive to Functional Form? pp. 1–28.

Sandler, D. and R. Sandler (2014). Multiple event studies in public finance and labor economics: A simulation study with applications. *Journal of Economic and Social Measurement 39*, 31–57.

Silberzahn, R. and many authors (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science 1*(3), 337–356.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics In Press*.

U.S. Census Bureau, Population Division (2020). Annual Estimates of the Resident Population for Incorporated Places in the United States: April 1, 2010 to July 1, 2019 (SUB-IP-EST2019-ANNRES), Release Date: May 2020. https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html. Accessed: 2020-07-02.

Villas-Boas, S. B., J. Sears, M. Villas-Boas, and V. Villas-Boas (2020). Are we# stayinghome to flatten the curve? *conditionally accepted in the American Journal of Health Economics*.

Weill, J. A., M. Stigler, O. Deschenes, and M. R. Springborn (2020). Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences 117*(33), 19658–19660.

# Appendices

# Appendix E

# Additional descriptive statistics

Table E.1: Summary Statistics for Main Control Variables

| Variable | N | Counties | Mean | Min | Max |
|---|---|---|---|---|---|
| Cases | 285417 | 3069 | 2.15 | 0.00 | 2174.00 |
| Cases > 0 | 285417 | 3069 | 0.27 | 0.00 | 1.00 |
| Cases sqrt | 285417 | 3069 | 0.34 | 0.00 | 46.63 |
| Deaths > 0 | 285417 | 3069 | 0.08 | 0.00 | 1.00 |
| Precipitation | 282997 | 3043 | 0.11 | 0.00 | 5.16 |
| Snow | 282997 | 3043 | 0.09 | 0.00 | 16.59 |
| Temperature (mean) | 282997 | 3043 | 43.80 | -18.39 | 87.75 |

Figure E.1: Timing of County Emergency Declarations in the U.S. from February 14 to June 11, 2020.

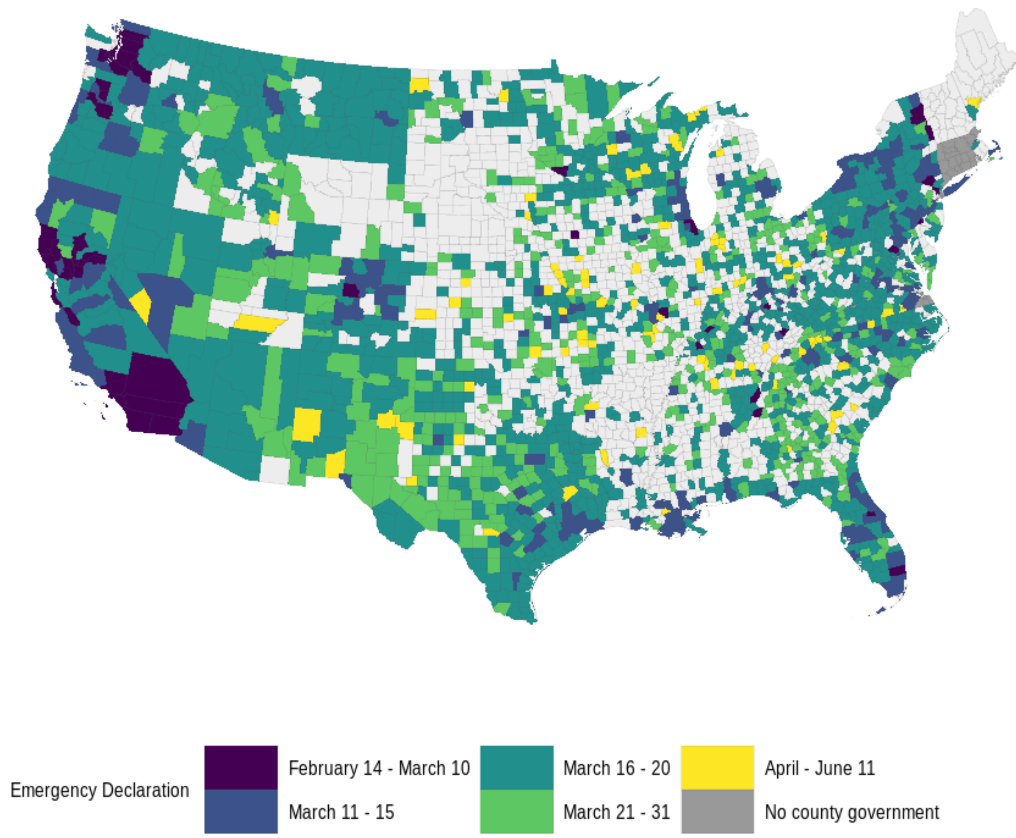
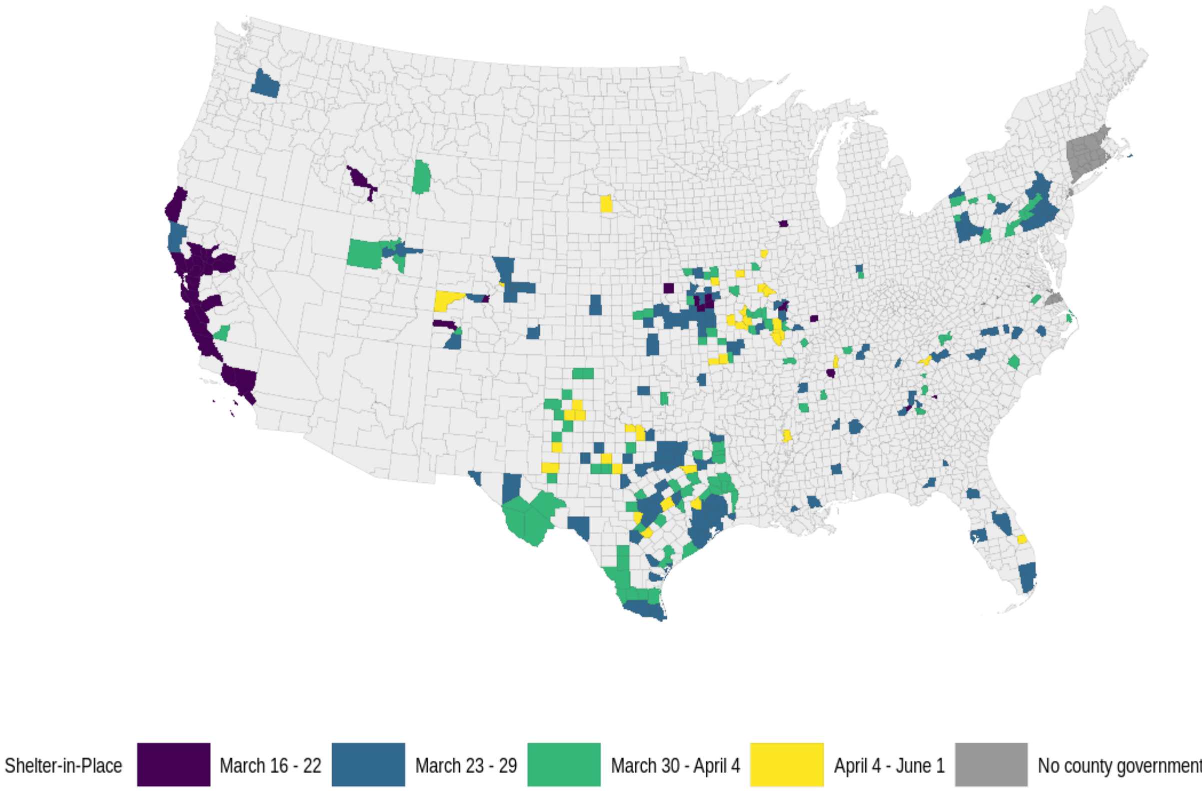


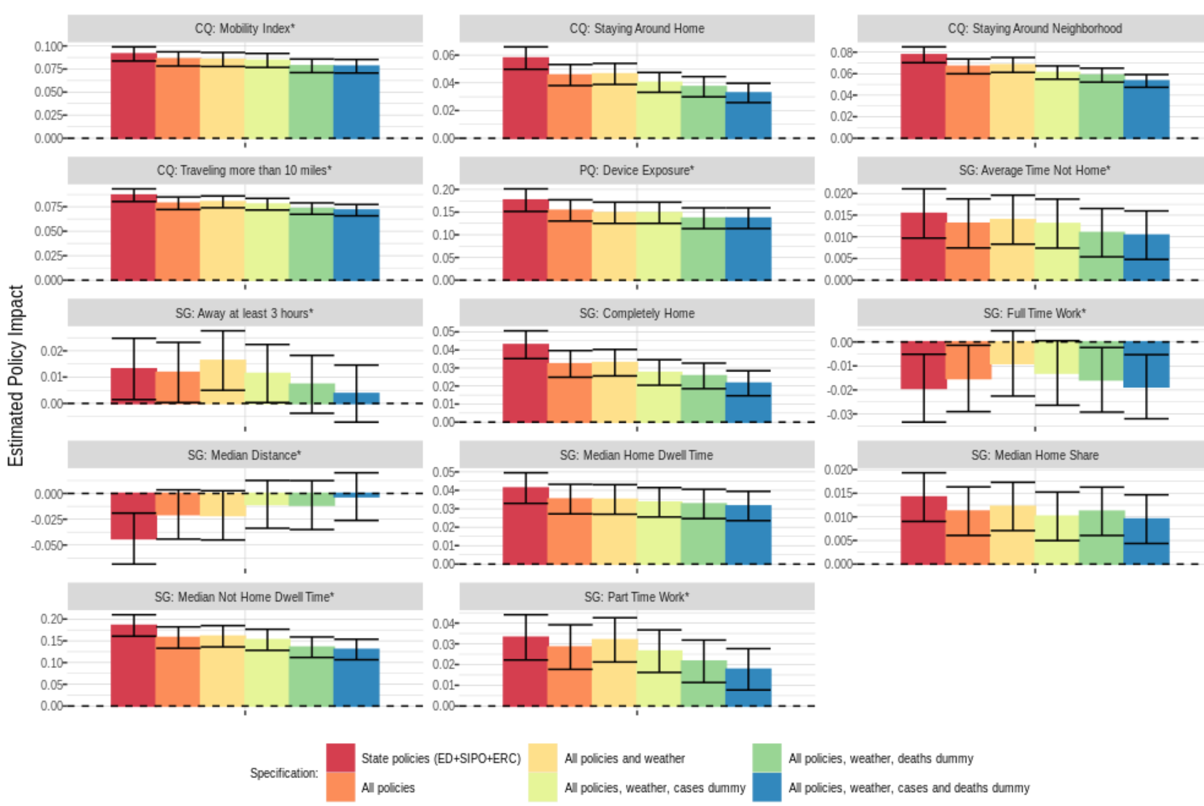
Figure E.2: Timing of County Shelter-In-Place Orders in the U.S. from February 14 to June 1, 2020



# Appendix F

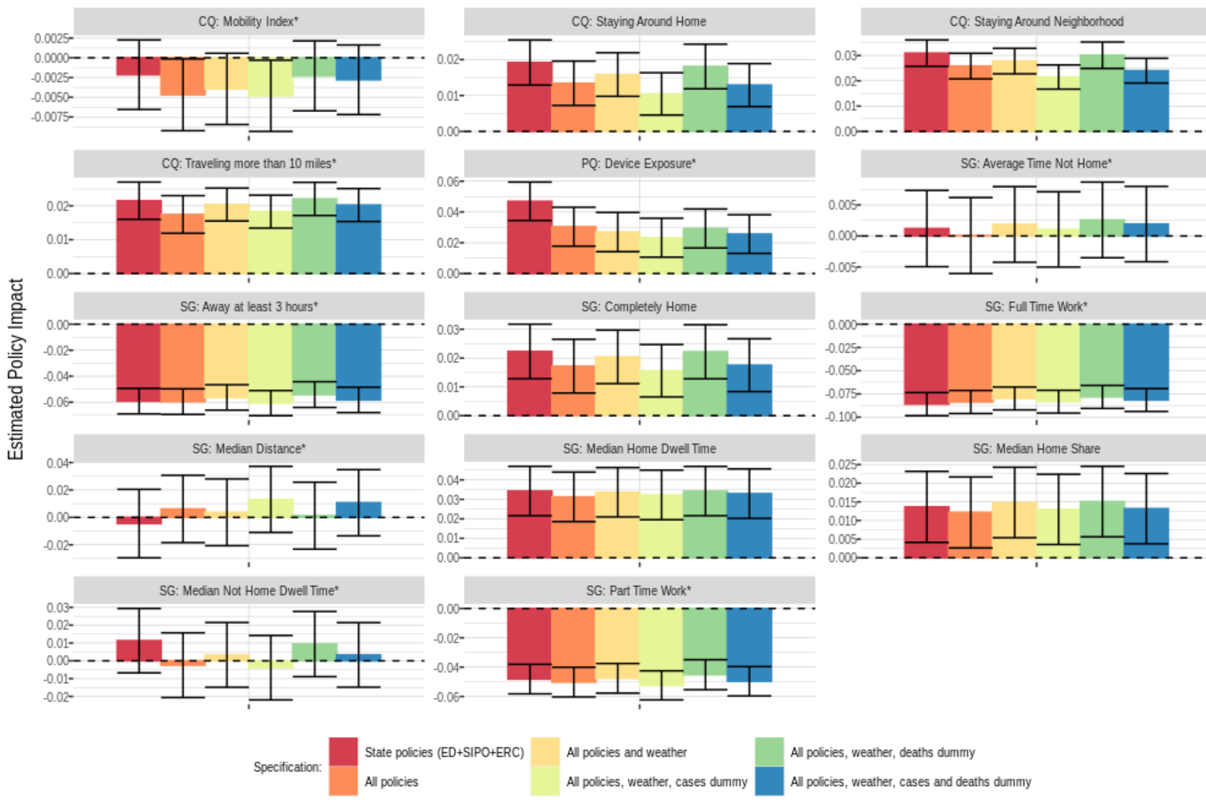
# Additional estimates of the impacts of MRPs on mobility

Figure F.1: Estimated Impact of State SIPO on Various Log Mobility and Activity Outcomes



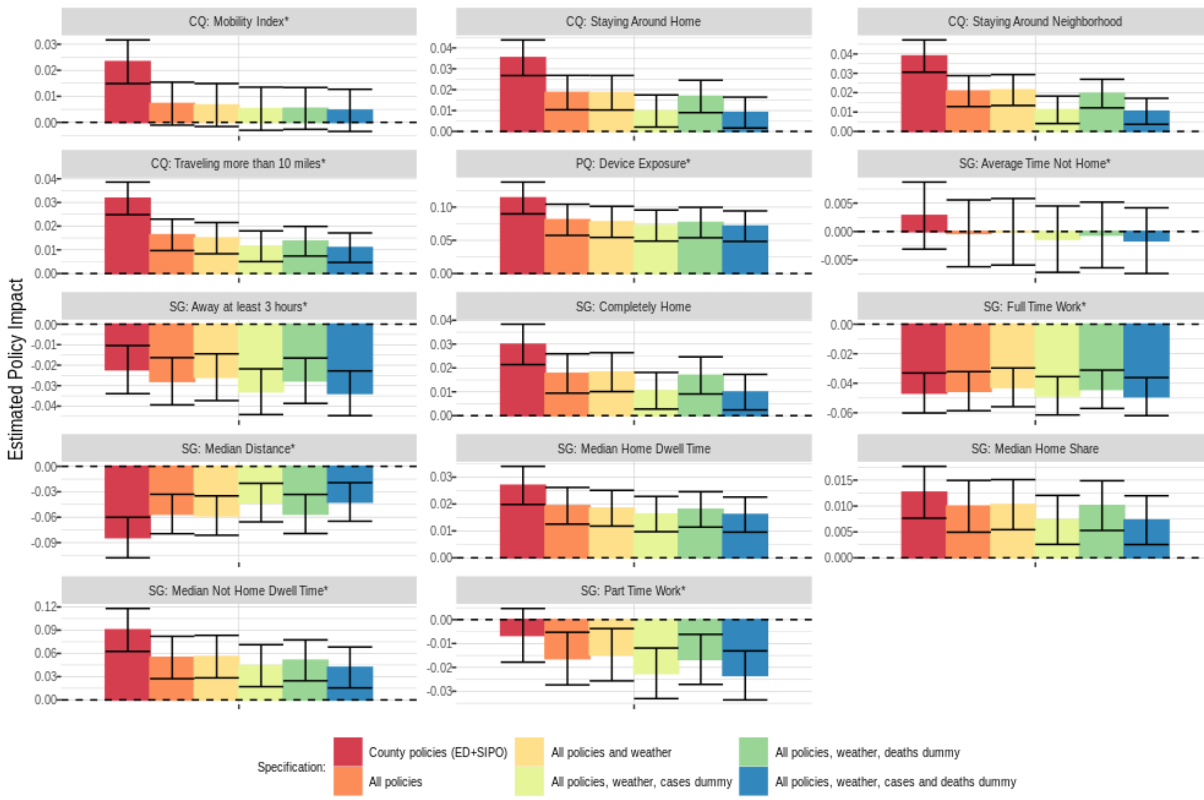

Notes: Figure A.1 reports the TWFE estimates of the impact of State SIPO on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.2: Estimated Impact of State ERC on Various Log Mobility and Activity Outcomes



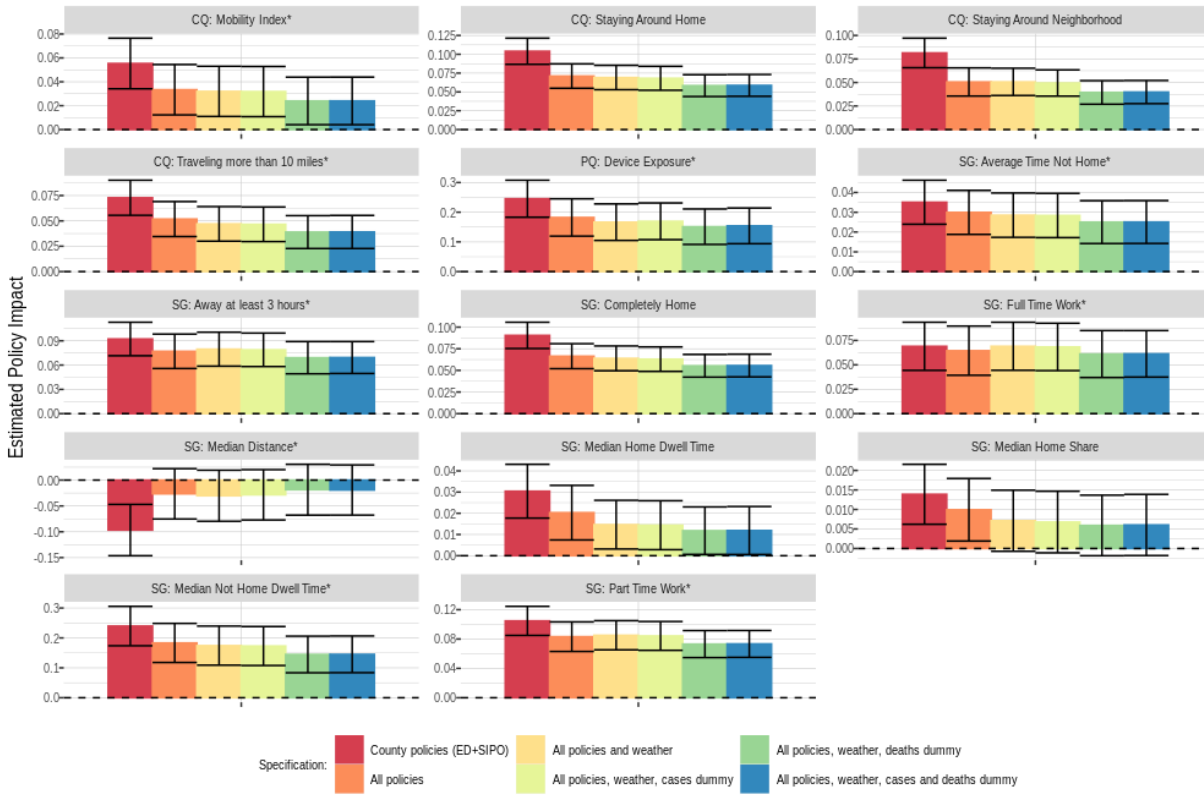

Notes: Figure A.2 reports the TWFE estimates of the impact of State ERC on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.3: Estimated Impact of County ED on Various Log Mobility and Activity Outcomes



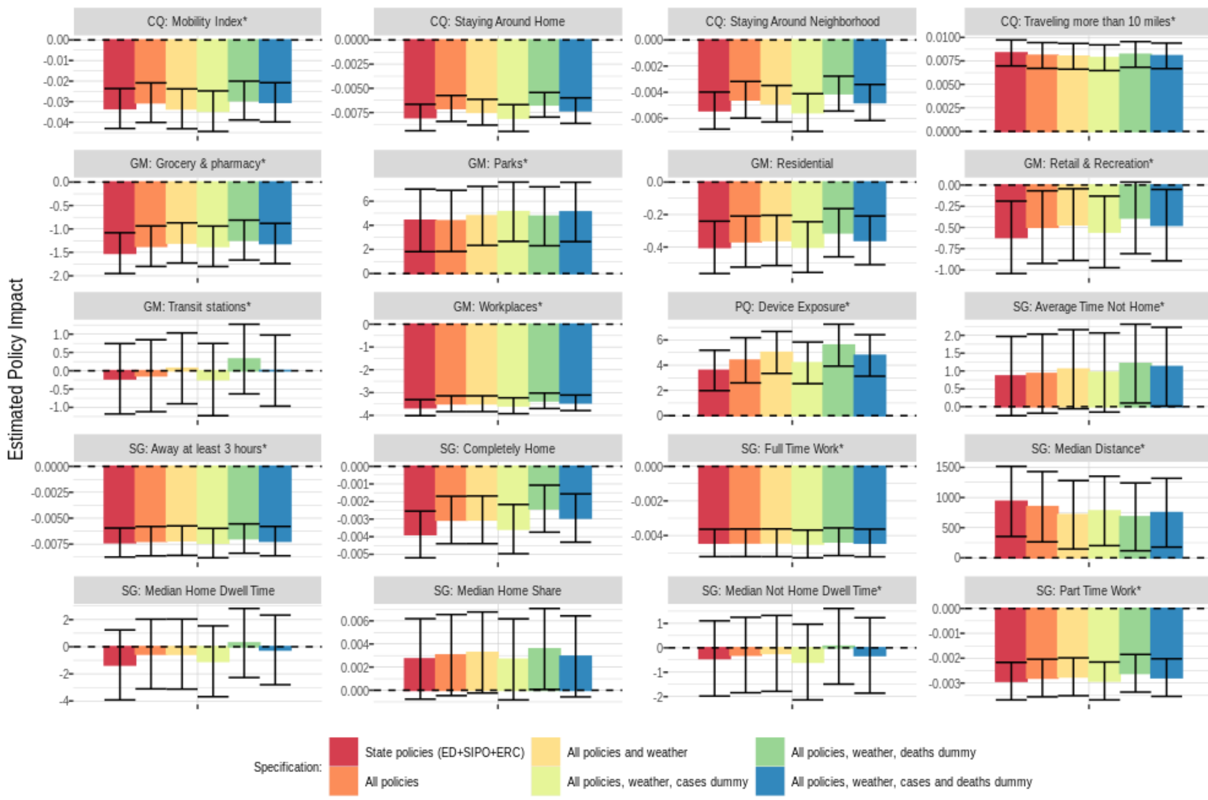

Notes: Figure A.3 reports the TWFE estimates of the impact of County ED on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.4: Estimated Impact of County SIPO on Various Log Mobility and Activity Outcomes



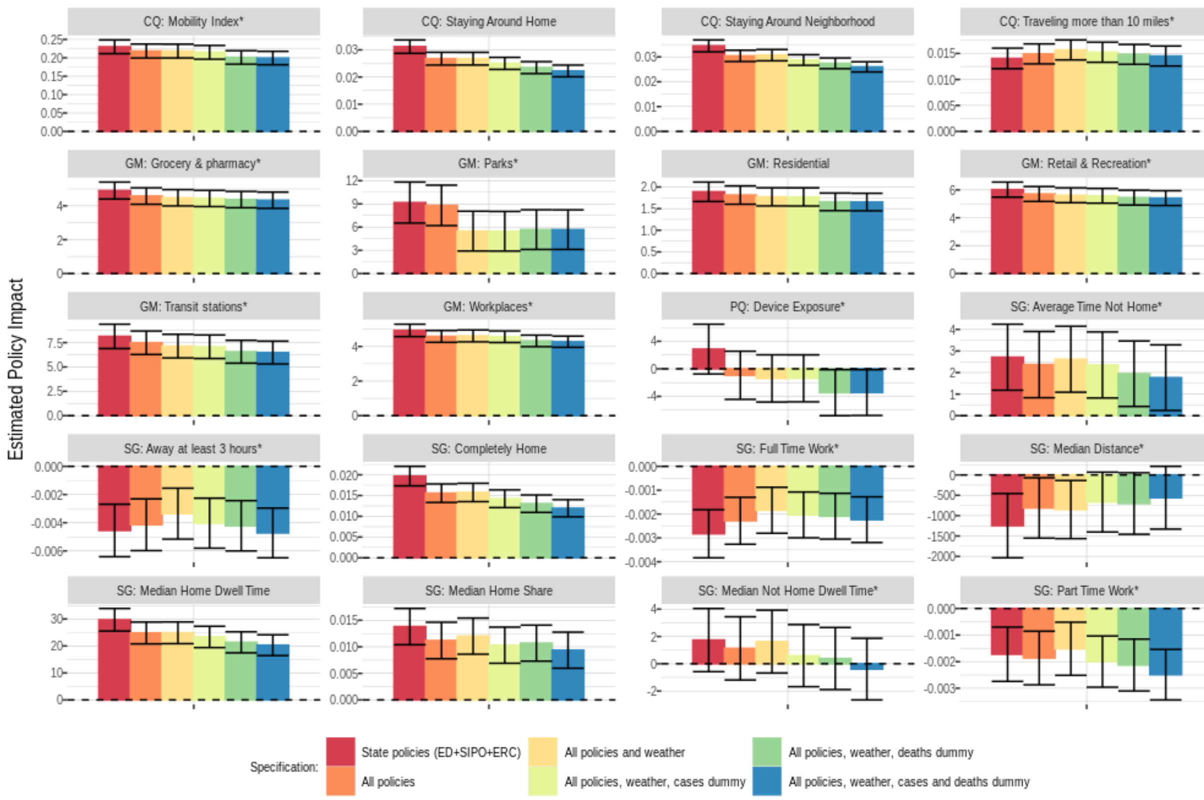

Notes: Figure A.4 reports the TWFE estimates of the impact of County SIPO on 14 mobility outcomes using a log-linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.5: Estimated Impact of State ED on Mobility and Activity Outcomes (in levels)



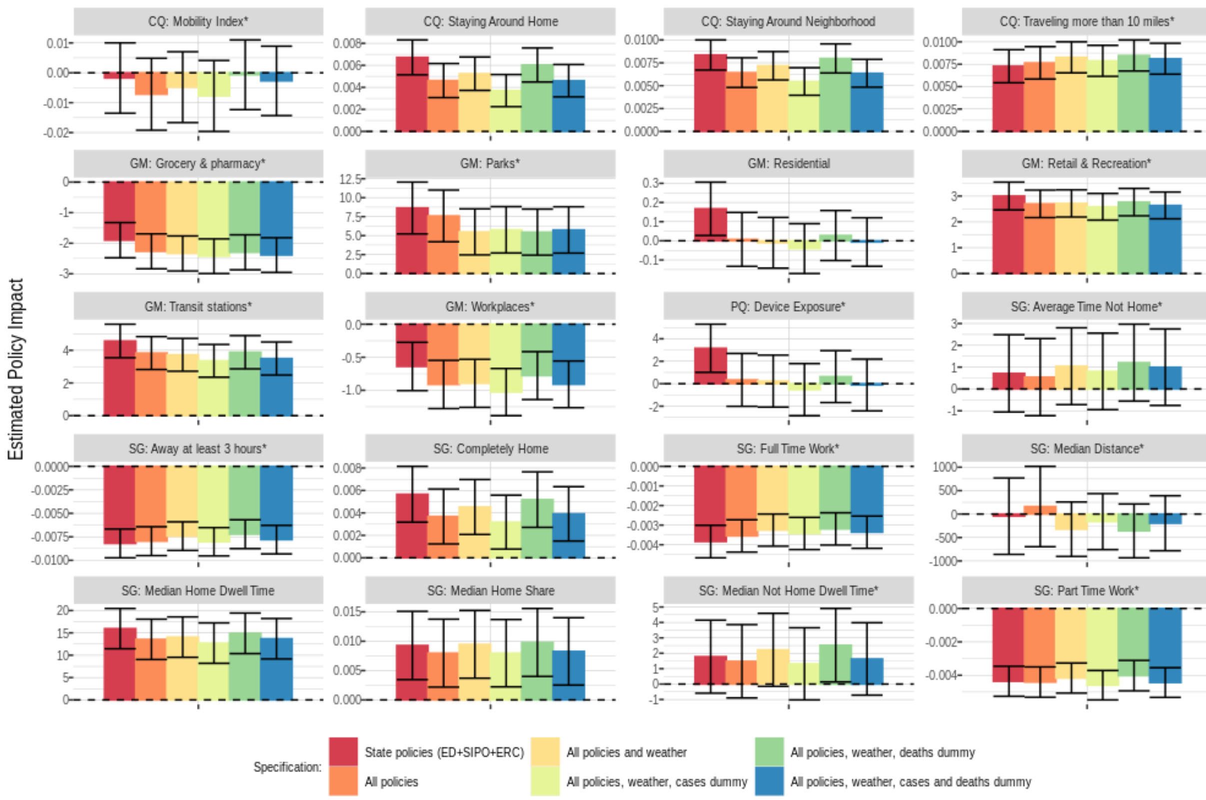

Notes: Figure A.5 reports the TWFE estimates of the impact of State ED on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.6: Estimated Impact of State SIPO on Mobility and Activity Outcomes (in levels)



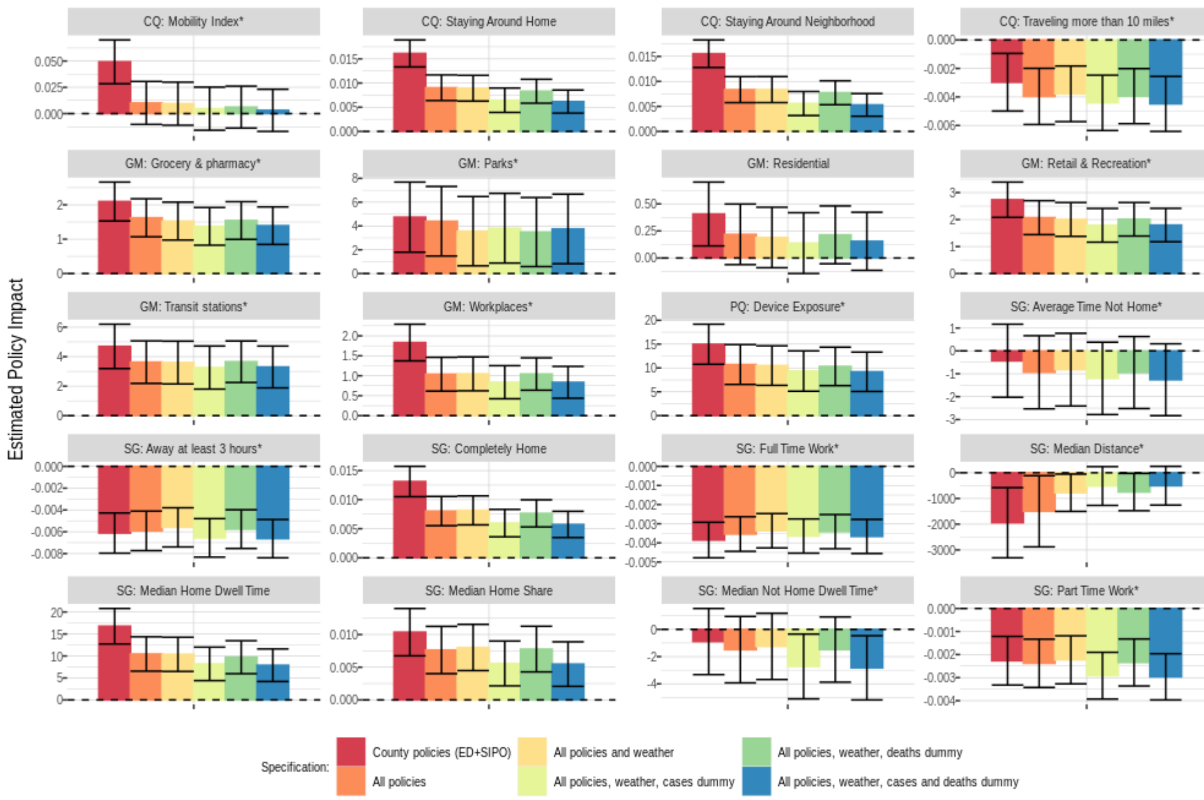

Notes: Figure A.6 reports the TWFE estimates of the impact of State SIPO on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.7: Estimated Impact of State ERC on Mobility and Activity Outcomes (in levels)



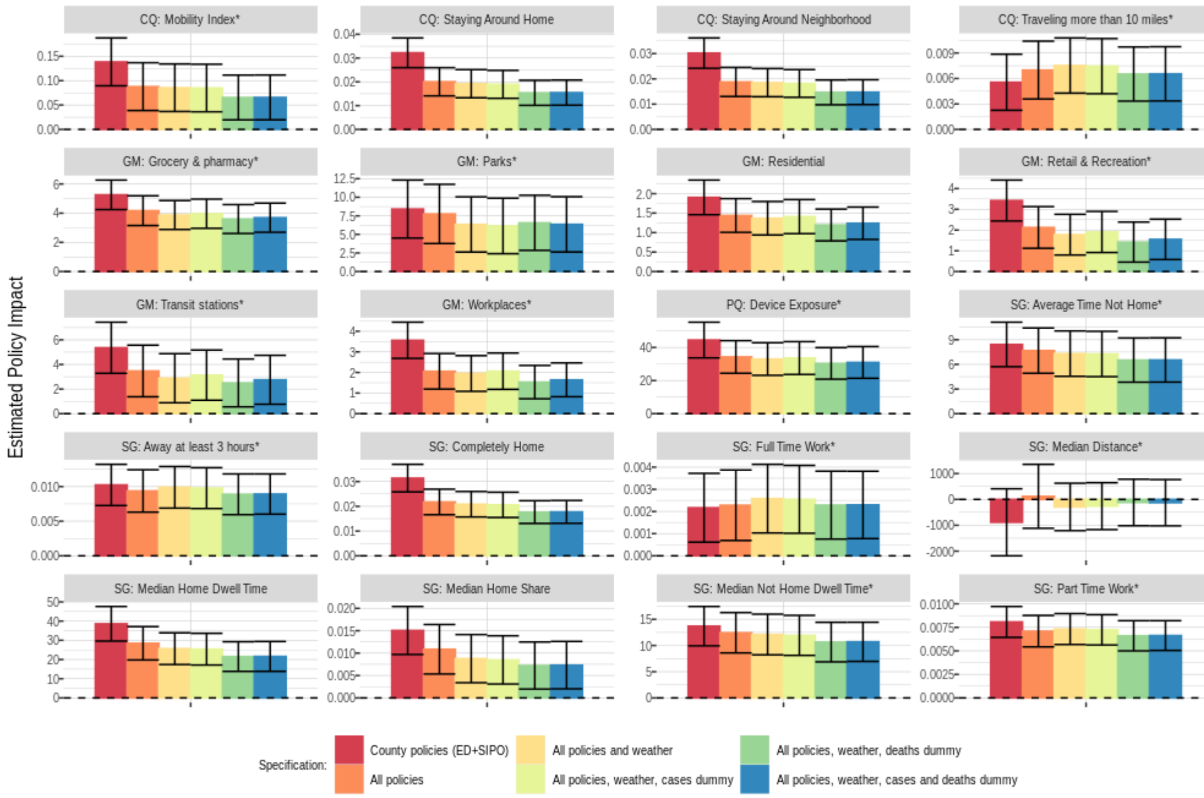

Notes: Figure A.7 reports the TWFE estimates of the impact of State ERC on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.8: Estimated Impact of County ED on Mobility and Activity Outcomes (in levels)



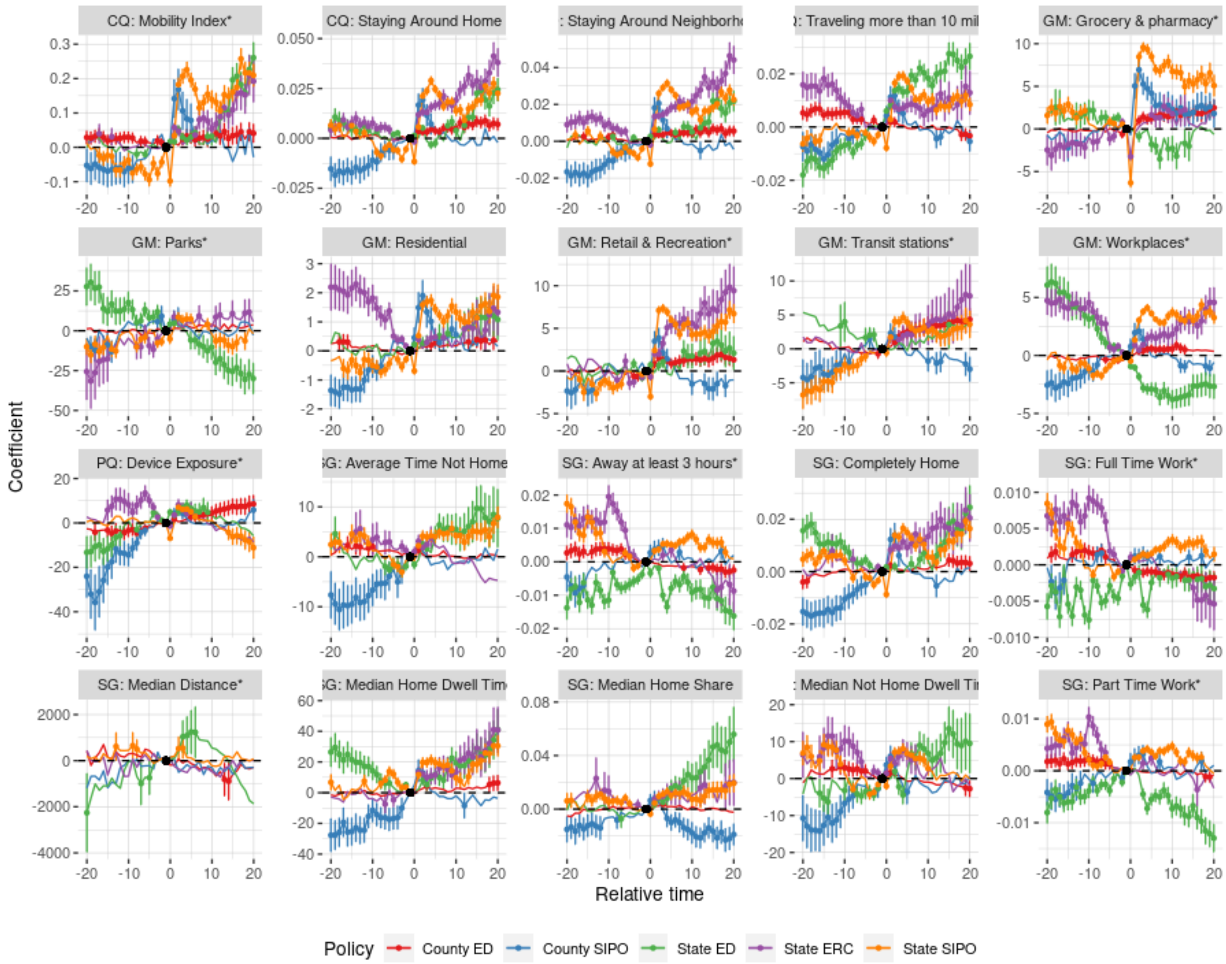

Notes: Figure A.8 reports the TWFE estimates of the impact of County ED on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.9: Estimated Impact of County SIPO on Mobility and Activity Outcomes (in levels)



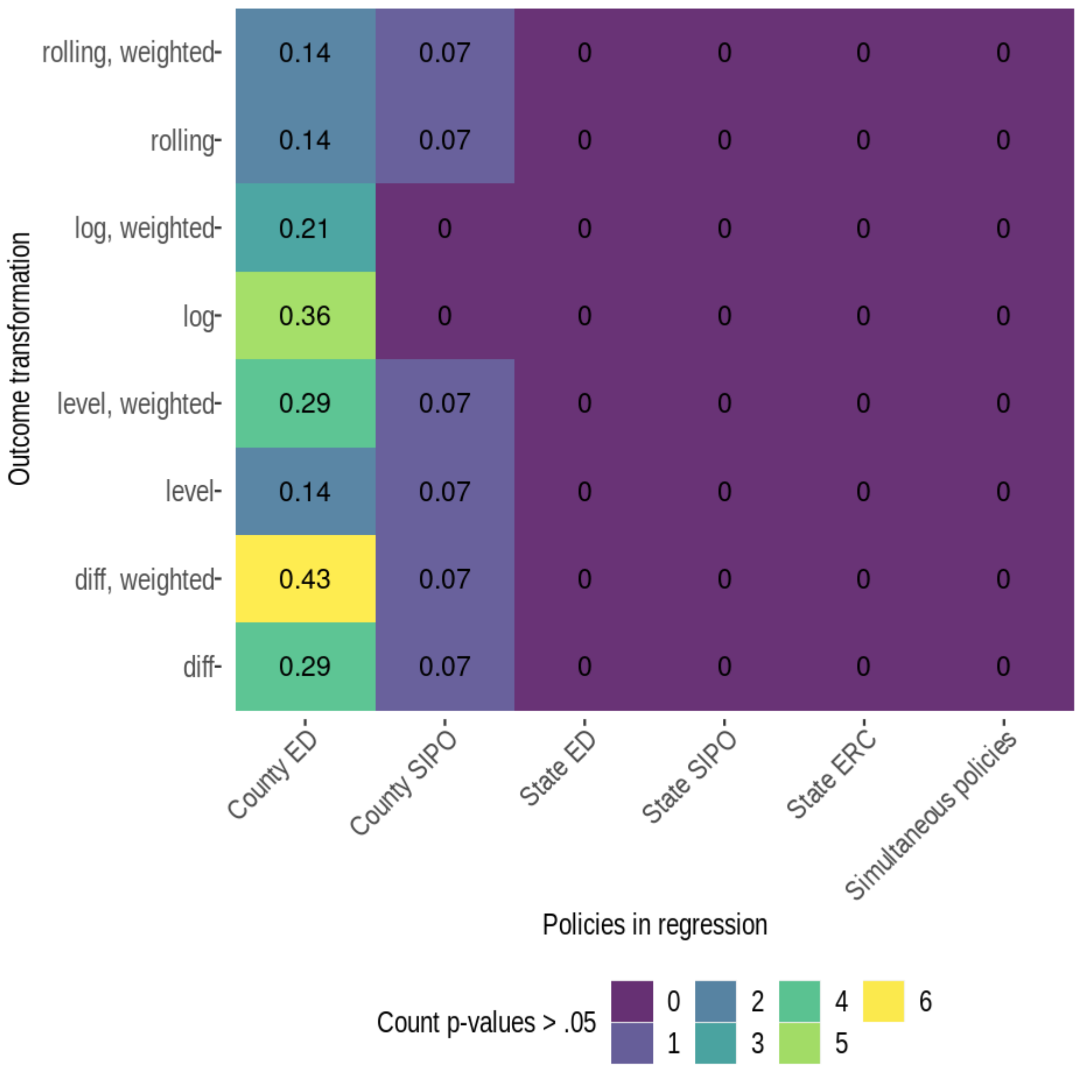

Notes: Figure A.9 reports the TWFE estimates of the impact of County SIPO on 20 mobility outcomes using a linear model. The color of the bars corresponds to the specification of the model (what MRPs and controls are included). The height of each bar represents the magnitude of the point estimate while the whiskers represent the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.10: Estimated Event-Study Coefficients for Mobility and Activity Outcomes (in levels)




Notes: Figure A.10 reports the TWFE estimates of the event-study coefficients as specified in Equation (2). MRPs under consideration are represented by the color of each line. The whiskers show the 95% confidence intervals based on standard errors clustered at the county-level.

Figure F.11: Heatmap of F-tests for Pre-Trend Tests in Event-Study Regressions




Notes: Figure A.11 is a heat map based on the p-values from F-tests testing the null hypothesis that the pre-MRP-adoption coefficients are jointly equal to zero in Event-study regressions. Each cell summarizes the results of 14 F-tests on pre-trends coefficients, corresponding to the 14 non-negative mobility outcomes from Safegraph, Cuebiq, and PlaceIQ (Google Mobility outcomes are ignored here since we cannot apply the log transformation on them). The number in each cell corresponds to the fraction of tests (out of 14) where the p-value on the null hypothesis exceeds 0.05, indicating that the null hypothesis would be rejected at the usual 5% significance level. The rows indicate the transformation of the dependent variable while the columns indicate the MRP considered (estimated individually in columns 1-5, and estimated jointly in column 6).

Figure F.12: Callaway and Sant'Anna Estimates of MRP Impacts by Treatment Cohort

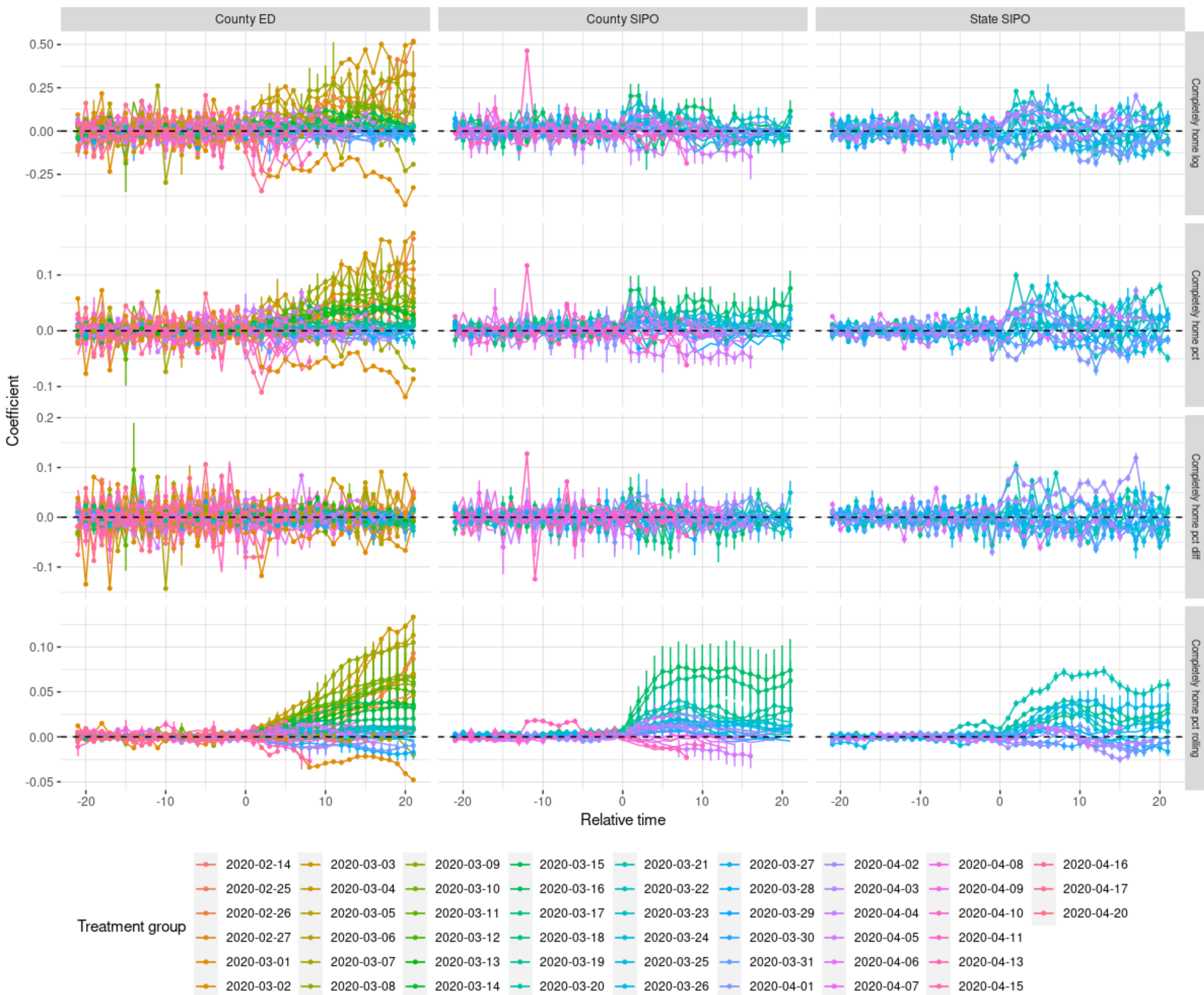
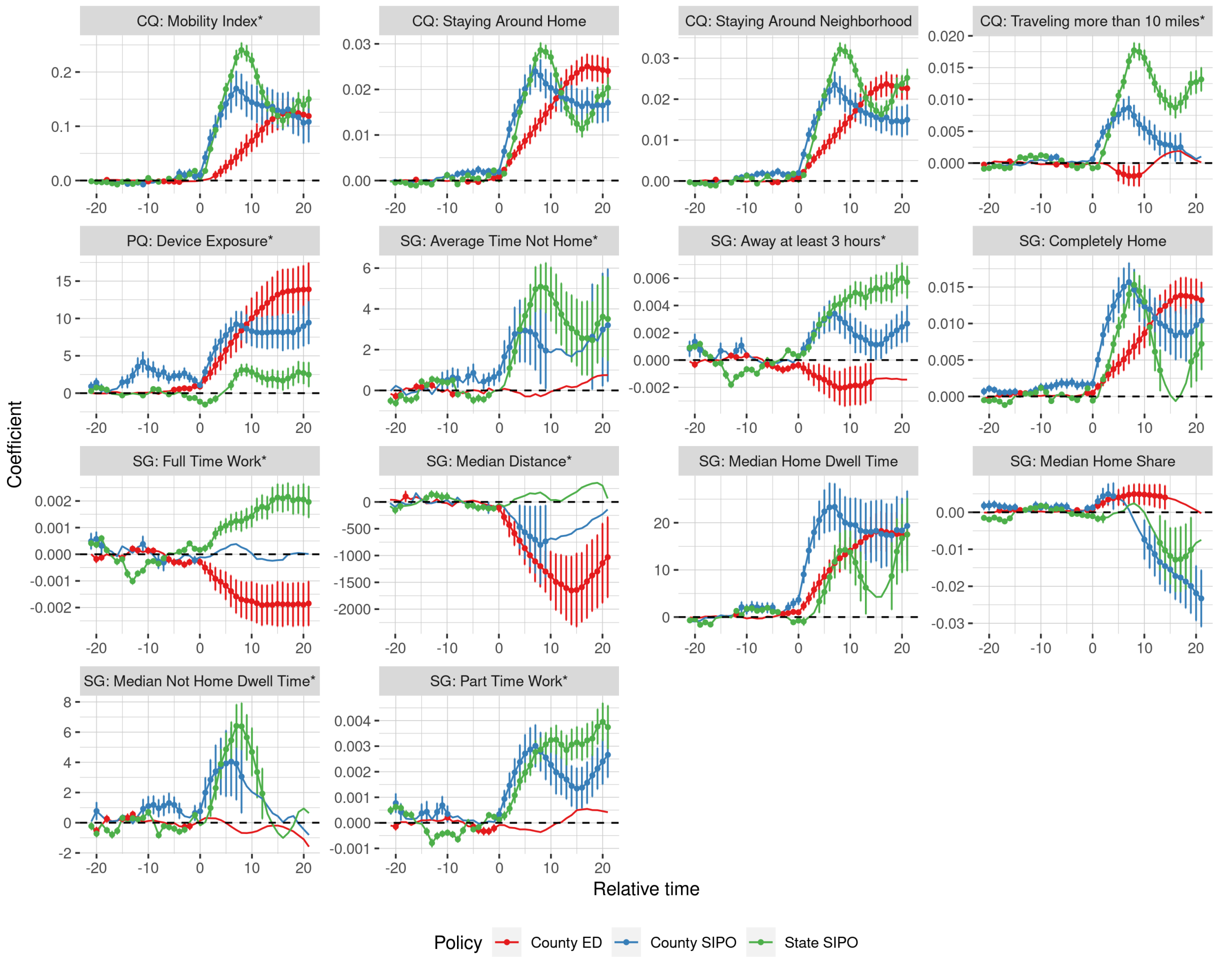
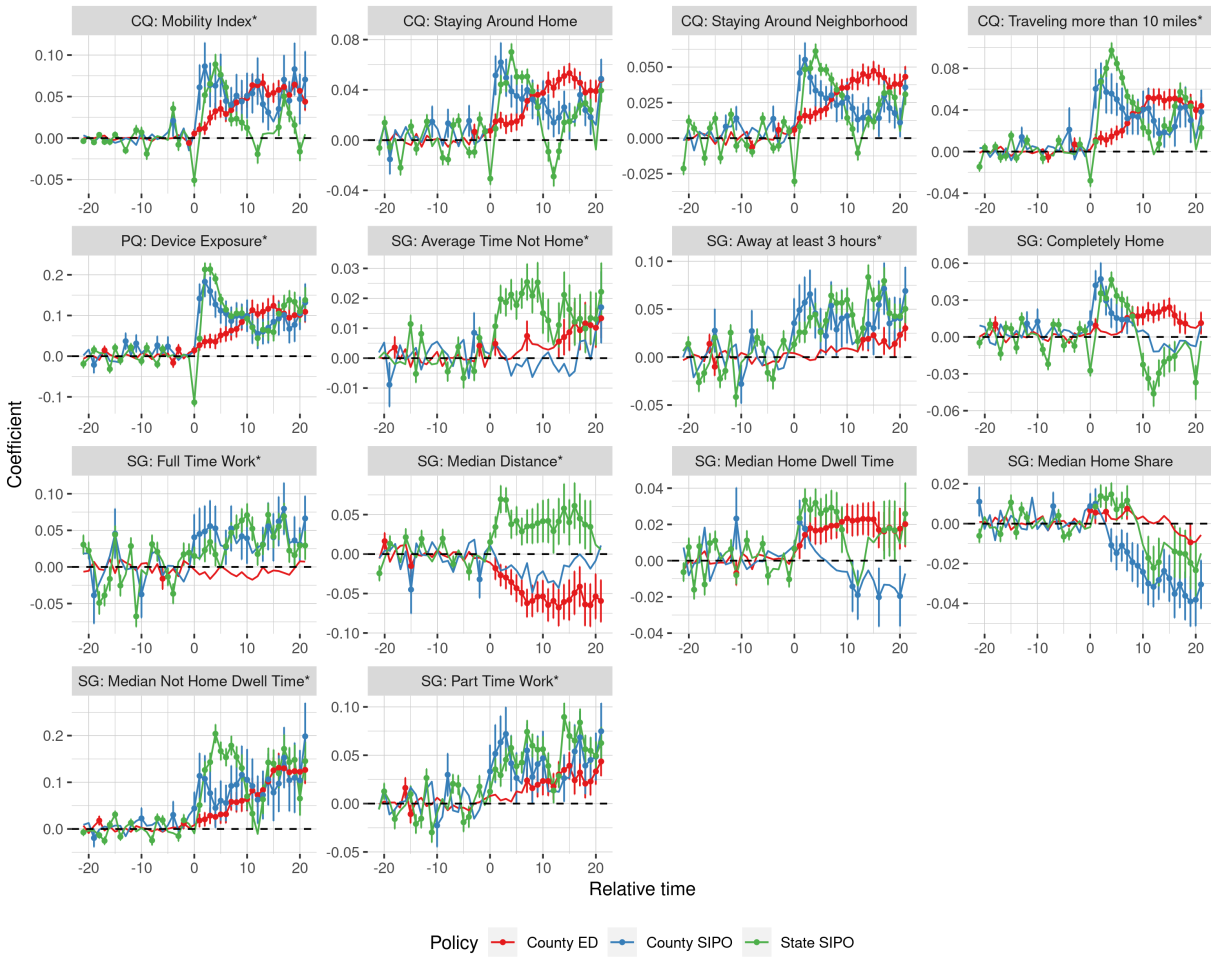


Figure F.13: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes Transformed with a 7-Day Moving Average



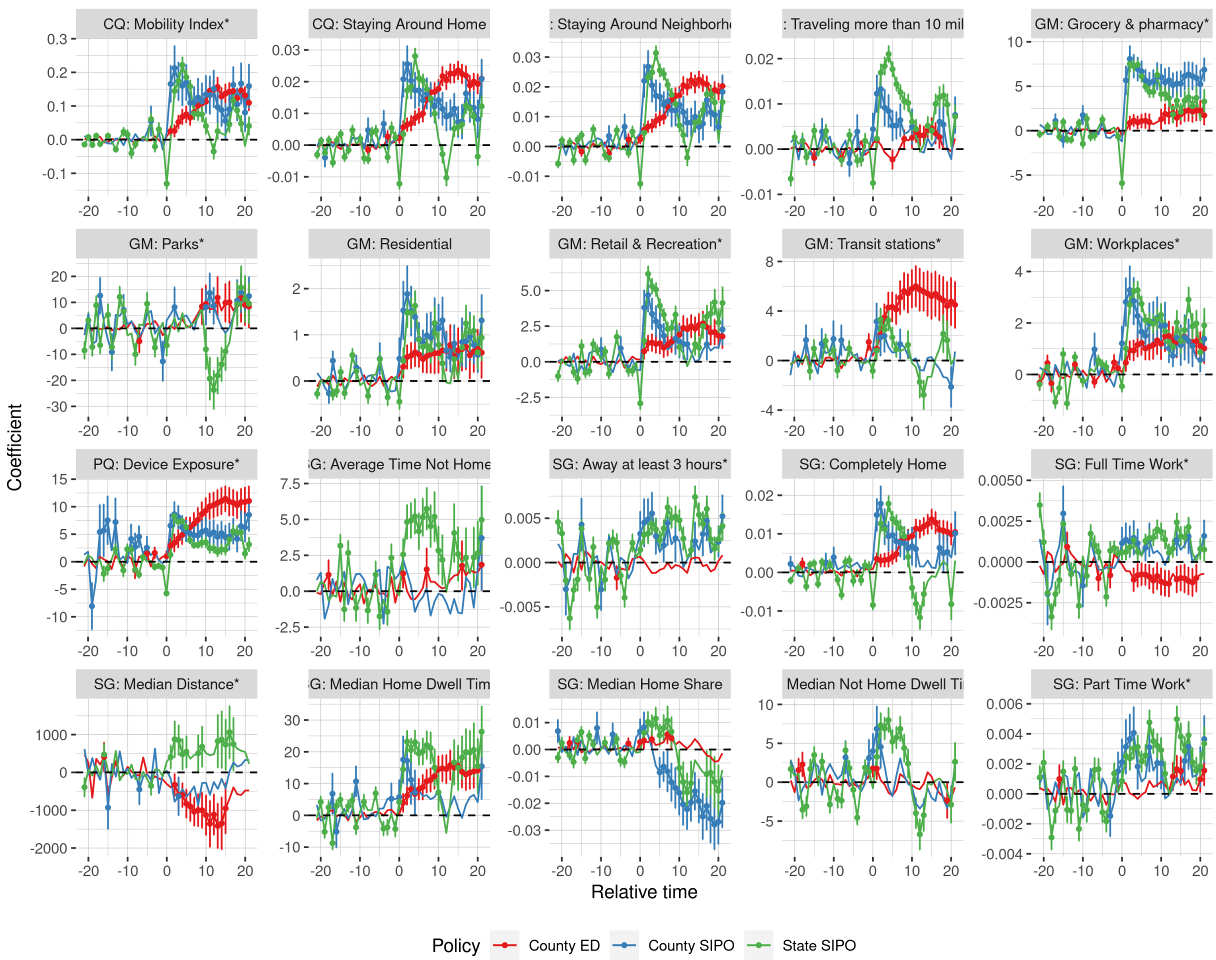

Notes: Figure A.12 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure F.14: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes in Logs




Notes: Figure A.13 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure F.15: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes Transformed with a First-Difference




Notes: Figure A.14 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

Figure F.16: Callaway and Sant'Anna Estimates of MRP Impacts Aggregated Across Treatment Cohorts, with Outcomes in Levels


Notes: Figure A.15 reports heterogeneity-robust estimates of the ATT, using the method in Callaway and Sant'Anna (2020). The MRP impact estimates are obtained from models estimated separately by MRP and excluding covariates. The error bars represent 95% point-wise confidence intervals using the multiplier bootstrap.

# Appendix G

# Variables description

Table G.1: Details about the Mobility Indicators Used in the Paper

| **Variable** | **Source** | **Description** | **Units** |
|---|---|---|---|
| Cuebiq Mobility Index | Cuebiq | Calculated using a derivative factor indicating the distance between opposite corners of a box drawn around the locations observed for users on each day. The index for each county is the median of the aggregated movements of all users within a county. Values can be interpreted as: 5 - 100 km, 4 - 10 km, 3 - 1 km,2 - 100m,1 - 10m. An index of 2.5 for a county, would mean the median user in that county is traveling 250m | No unit |
| Staying Around Home | Cuebiq | Percentage of users staying at home in any given state/county. It is calculated by measuring how many users moved less than 330 feet from home | % |
| Staying Around Neighbor-hood | Cuebiq | Percentage of users traveling less than one mile from home | % |
| Traveling more than ten miles | Cuebiq | Percentage of users traveling more than ten miles from home | % |

Table G.1: Details about the Mobility Indicators Used in the Paper

| **Variable** | **Source** | **Description** | **Units** |
|---|---|---|---|
| Grocery and Pharmacy | Google Mobility | Changes in the number of visits to grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies, relative to baseline. The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020 | % |
| Parks | Google Mobility | Changes in the number of visits to places like national parks,public beaches, marinas, dog parks, plazas, and public gardens, relative to baseline. | % |
| Retail and Recreation | Google Mobility | Changes in the number of visits to places like restaurants,cafes, shopping centers, theme parks,museums, libraries, and movie theaters, relative to baseline | % |
| Residential | Google Mobility | Changes in the number of visits to places of residence relative to baseline. | % |
| Transit Stations | Google Mobility | Changes in the number of visits places like public transport hubs such as subway, bus, and train stations, relative to baseline | % |
| Workplaces | Google Mobility | Changes in the number of visits to places of work relative to baseline | % |
| Average Time Not Home | Safegraph | Average time a device is recorded as away from home within a county, reconstructed from ”bucketed” time away from home | min- utes |
| Full Time Work | Safegraph | County share of devices that spent greater than 6 hours at a location other than their home geohash-7 during the period of 8 am - 6 pm in local time | % |

Table G.1: Details about the Mobility Indicators Used in the Paper

| **Variable** | **Source** | **Description** | **Units** |
|---|---|---|---|
| Median Distance | Safegraph | Median distance traveled from the geohash-7 of the home by the devices during the time period (excluding any distances of 0). The median is provided at the census block group level and we take the weighted mean of these medians at the county level (using the number of devices as weights) | me-ters |
| Median Not Home Dwell Time | Safegraph | Median dwell time at places outside of geohash-7 home for all observed devices during the time period. For each device, the observed minutes outside of home across the day (whether or not these were contiguous) are summed to get the total minutes for each device. The median is first calculated across all these devices at the census block group level, and then at the county level | min-utes |
| Part Time Work | Safegraph | County share of devices that spent one period of between 3 and 6 hours at one location other than their geohash-7 home during the period of 8 am - 6 pm in local time. This does not include any device that spent 6 or more hours at a location other than home | % |
| Away at least three hours | Safegraph | County share of devices that spent at least 3 hours at one location other than their geohash-7 home during the period of 8 am - 6 pm in local time | % |
| Completely Home | Safegraph | County share of devices which did not leave the geohash-7 in which their home is located during the time period. | % |
| Median Home Dwell Time | Safegraph | Median dwell time at home geohash-7 ("home") in minutes for all devices during the time period. For each device, the observed minutes at home across the day are summed (whether or not these were contiguous) to get the total minutes for each device. The median is provided at the census block group level and we take the weighted mean of these medians at the county level | min-utes |

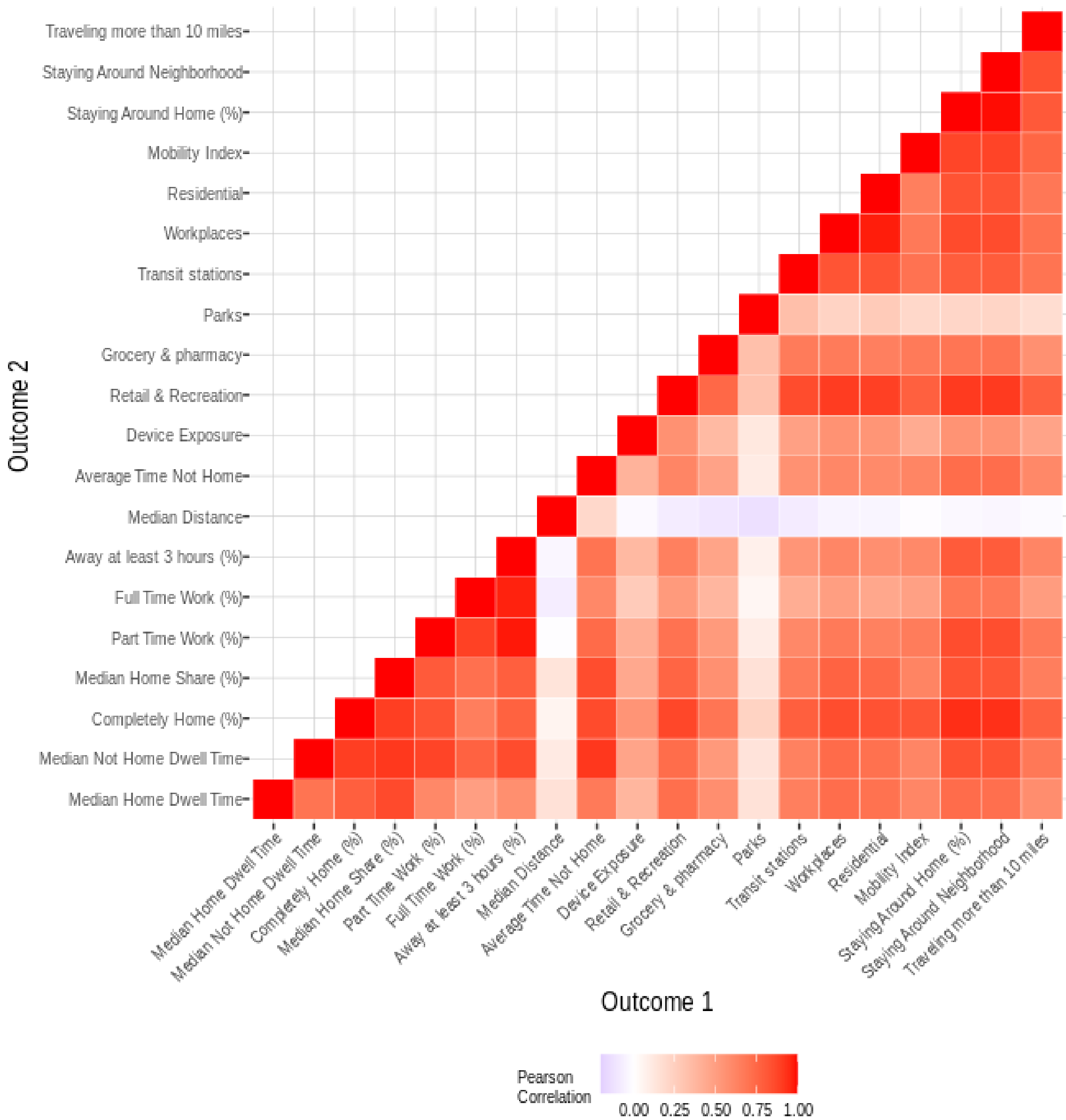
Table G.1: Details about the Mobility Indicators Used in the Paper

| **Variable** | **Source** | **Description** | **Units** |
|---|---|---|---|
| Median Home Share | Safegraph | Median percentage of time we observed devices home versus observed at all during the time period. The median is provided at the census block group level and we take the weighted mean of these medians at the county level | % |
| Device Exposure | PlaceIQ | For a device, number of distinct devices that also visited any of the commercial venues that this device visited that day. The county-level DEX reports the county-level average of this number across all devices residing in the county that day. We use the adjusted version of this variable, which accounts for devices not seen living their home (further details provided by the authors. | Count |

# Appendix H

# Correlation between mobility measures

Figure H.1: Pearson correlations between mobility outcome variables



# Appendix I

# Results from Bacon-Goodman decomposition

The recent econometric literature highlights that two-way fixed effect estimates do not generally identify an average treatment effect when treatment is staggered over time (de Chaisemartin and D'Haultfœuille, 2020b; Goodman-Bacon, 2021). In particular, Goodman-Bacon (2021) shows that the two-way fixed effect estimator can be decomposed as a weighted sum of all possible 2x2 difference-in-differences estimators generated by a staggered policy adoption design. Formally, let $k = 1...K$ be different groups of units ordered by the time at which they receive a binary treatment, and U is a group that never receives treatment. Let $\bar{y}_b^{POST(a)}$ denote the sample mean of $y_{it}$ in group $b$ during group $a$ post period, and define $\bar{y}_b^{PRE(a)}$ similarly. Finally, let $\bar{y}_a^{MID(a,b)}$ be the sample average of units in group $a$ after group $a$ becomes treated, but before group $b$ becomes treated. Then Theorem 1 of Goodman-Bacon (2021) shows that in the two-way fixed effect model with a unique binary policy and no covariates, the DD estimator can be written as:

$$\widehat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\beta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{\ell > k} \left[ s_{k\ell}^{k} \widehat{\beta}_{k\ell}^{2x2,k} + s_{k\ell}^{\ell} \widehat{\beta}_{k\ell}^{2x2,\ell} \right]$$

where the $2 \times 2$ diff-in-diffs estimators are:

$$\begin{aligned}
\widehat{\beta}_{kU}^{2x2} &\equiv \left( \bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)} \right) - \left( \bar{y}_U^{POST(j)} - \bar{y}_U^{PRE(j)} \right) \\
\widehat{\beta}_{k\ell}^{2x2,k} &\equiv \left( \bar{y}_k^{MID(k,\ell)} - \bar{y}_k^{PRE(k)} \right) - \left( \bar{y}_\ell^{MID(k,\ell)} - \bar{y}_\ell^{PRE(k)} \right) \\
\widehat{\beta}_{k\ell}^{2x2,\ell} &\equiv \left( \bar{y}_\ell^{POST(\ell)} - \bar{y}_\ell^{MID(k,\ell)} \right) - \left( \bar{y}_k^{POST(\ell)} - \bar{y}_k^{MID(k,\ell)} \right)
\end{aligned}$$

and the weights $s_{kj}$ are functions of group sizes and variance of treatment within groups.

This decomposition highlights two important potential pitfalls of the TWFE estimator. First, due to the weights, any amount of treatment effect heterogeneity between groups means that the $\beta^{\hat{D}D}$ estimator can be quite different from the average treatment effect on the treated. Furthermore, the $\hat{\beta}^{DD}$ estimator implicitly compares units that are treated early to units that are treated "later" (the $\hat{\beta}^{2x2,\ell}_{kl}$ components). Thus if the treatment effect is time-varying, these comparisons will not be informative about the (time-varying) average treatment effect on the treated.

To assess whether identification issues related to treatment effect heterogeneity in bias our baseline TWFE estimates, we first estimate a modified version of (3.1) above including each policy in isolation and no covariates. We then implement the Bacon-Goodman decomposition above, but remove all $\hat{\beta}^{2x2,\ell}_{kl}$ components and normalize the remaining weights so that they sum to one. We finally compute the ratios of these two estimators: the standard TWFE and the adjusted TWFE that excludes the "treated earlier vs. later" comparisons.

$$\frac{\sum_{k\neq U} \widetilde{s}_{kU} \widehat{\beta}^{2x2}_{kU} + \sum_{k\neq U} \sum_{\ell>k} \left[ \widetilde{s}^{k}_{k\ell} \widehat{\beta}^{2x2,k}_{k\ell} \right]}{(\sum_i X_i' X_i)^{-1} (\sum_i X_i' Y_i)}$$

Figure I.1: Ratio of weighted coefficients with "Later vs Earlier" comparisons removed to regular 2-way fixed effects coefficients

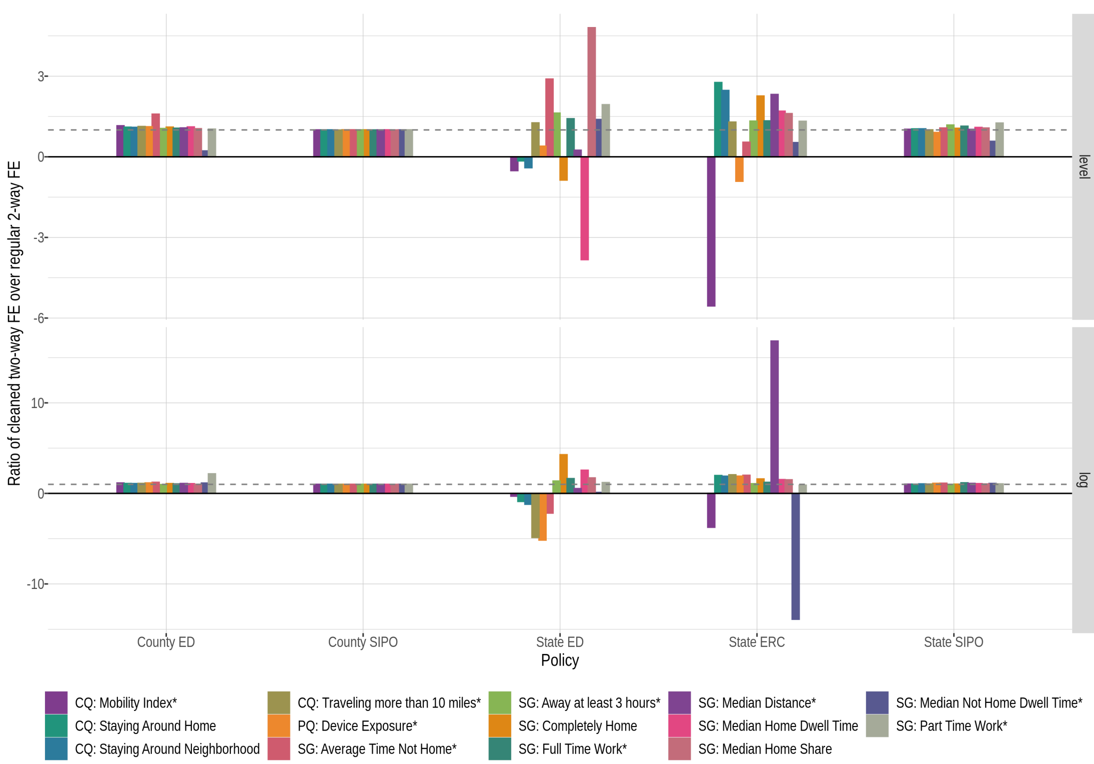


Figure I.1 presents the ratios for 14 outcome variables in levels and in logs (Google Mobility outcomes are excluded since they include a lot of zeros). We note that for the county policies and the State Shelter-in-place orders, the ratios are very close to one for most outcomes: removing the "treated later vs treated earlier" comparisons does not substantially impact our estimates. However, for the State ED and State ERC policies the discrepancy can be very large, with some estimates having opposite signs depending on the outcome variable considered. This result is somewhat expected from the Bacon Goodman decompositions: for State ED and State ERC there is no "true" control group to estimate the impact of the policy. All units are subject to a state ED and ERC at some point. As a result, removing the "treated later vs treated earlier" comparisons can have a large impact on the overall estimate. These discrepancies lead us to suspect dynamic treatment effects, which we further investigate with event studies in section 3.5.

# Chapter 4

# Housing Assistance, Neighborhood Quality, and Environmental Inequalities

**Joint work with John L. Voorheis. U.S. Census Bureau. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release, authorization number CBDRB-FY2021-CES010-025**

## 4.1 Introduction

More than 2.3 million low-income American families receive federal rental assistance in the form of housing choice vouchers (HCV, also called "Section 8 Vouchers"), administered by the Department of Housing and Urban Development (HUD). These vouchers subsidize rents for low and extremely low income households within the private market: recipient families usually pay 30% of their income in rent, while the federal government pays the rest up to a maximum amount.[1] Currently the nation's largest source of rental assistance, vouchers aim not only to improve housing security by reducing rent burden, but also to move disadvantaged families to wealthier and cleaner neighborhoods, in line with a large and growing literature that highlights causal impacts of neighborhoods and environmental quality on the long term health and economic outcomes of adults (Katz et al., 2001; Kling et al., 2007; Ludwig et al., 2013) and children (Chetty et al., 2016; Alexander and Currie, 2017; Chetty and Hendren, 2018; Isen et al., 2017; Colmer and Voorheis, 2020) – see Chyn and Katz (2021) for a recent literature review.

However, although housing vouchers were found to successfully reduce rent burden and housing insecurity, previous studies found that many voucher recipients still live in economically disadvantaged neighborhoods (Pendall, 2000; DeLuca et al., 2013; Holloway, 2014; Kleit et al., 2015; Ellen, 2020) For instance, Collinson and Ganong (2018) remark that "[...] housing voucher holders opt to live in neighborhoods of much lower quality than the average neighborhood, and typically live in neighborhoods similar to their neighborhood before receiving a voucher". Previous work found evidence of multiple constraints that prevent voucher recipients from moving to higher quality neighborhoods (DeLuca et al., 2013; DeLuca and Rosenblatt, 2017; Bergman et al., 2019). In particular, several papers find that landlords routinely discriminate against section-8 voucher recipients (Phillips, 2017; Rosen, 2020), sometimes openly advertising "No Section 8" on their listings (Tighe et al., 2017; Tobias and Botts, 2019; Versoulis, 2020).

In this paper, we re-examine the evidence on the impact of housing choice vouchers on residential outcomes by using restricted administrative data on the entire population of housing choice voucher users over the past two decades. We use new administrative records to construct the residential history of voucher users both *before and after* they start using federal vouchers, and pair these records with a unique dataset of economic and environmental metrics resolved at the neighborhood level. We have three main findings.

First, we find strong evidence of dynamic selection into housing choice voucher usage: future voucher users move to neighborhoods that are increasingly *poorer* until they use their first voucher. Such patterns imply that simple comparisons between "treated" individuals (who receive a voucher) and individuals who do not receive a voucher are unlikely to deliver credibly causal estimates of the impact of housing choice vouchers, even after controlling for or matching on observable characteristics.

We then use recently-developed difference-in-differences estimators to quantify the impacts of vouchers on residential outcomes by comparing the locations of beneficiaries who receive and use their first voucher at different times. We find that the effects exhibit sharp dynamics: housing choice vouchers initially cause families to move to neighborhoods that are of lower "quality" than the neighborhood they are starting from,

[1] Households are considered low-income if they earn 80% or less of the local median income, and extremely-low-income if they earn below the poverty line or 30% of the local median, whichever is higher (Center on Budget and Policy Priorities, 2020)

but the sign of the effect quickly reverses: after two to four years, depending on the race/ethnic group considered, the average effect of vouchers on neighborhood income becomes positive, with average annual increases from +1% to +1.5% in neighborhood income.

Finally, our results reveal substantial heterogeneity between states and neighborhood quality metrics. Vouchers cause long-term moves to higher income neighborhoods in California, Florida and Texas, but not in New York. Effects on environmental outcomes are mixed depending on the state and race/ethnic group considered: we estimate that vouchers cause small declines in air pollution exposure (as measured by the fine particulate matter concentration, or PM2.5) among all groups, while vouchers cause White families to be less exposed to toxic release inventory sites, an effect driven by migrations to more urban areas.

Taken together, our results suggest that housing choice vouchers do contribute to moving families to wealthier and cleaner neighborhoods in the long-run. Effects are similar between race and ethnic groups, but are strongly dynamic, with initial effects that are negative. Such dynamics are consistent with a "learning" mechanism, where families have to learn to navigate a complex administrative process in order to first obtain and then use the vouchers, as highlighted in recent empirical results from randomized experiments in Seattle and King County, WA (Bergman et al., 2019). These results also suggest that vouchers' expiration dates may be inducing households to use their vouchers in poorer neighborhoods in order to maintain their benefits. In ongoing work, we attempt to disentangle these mechanisms.

The rest of this paper proceeds as follows. Section 4.2 presents the data as well as several descriptive statistics. Section 4.3 provides descriptive evidence on dynamic selection and treatment effects, which informs our empirical approach presented in section 4.4. Section 4.5 concludes.

## 4.2 Data and descriptive statistics

We consolidate restricted-access data from the Department of Housing and Urban Development (HUD) on all individuals who used a housing choice voucher (HCV) at any point between 2000 and 2018. These data contain information on family members, housing characteristics, and assistance payment amounts. Because we are interested in the impacts of vouchers on the residential outcomes of families and children, we focus on women who are the head of household, or on the head of household's partner when the head of household is a man. This further ensures that for each family, movements between neighborhoods are only counted once.

While the consolidated HUD files provide comprehensive information on the universe of voucher users, they suffer from two important limitations. First, these files only consistently record individuals who receive *and* use a housing choice voucher. Individuals who apply to receive a housing choice voucher but are not issued one are not recorded, while individuals who are issued a voucher but do not use it are not consistently recorded in the dataset.[2] As a result, the current analysis is restricted to the universe of vouchers *users*. In the remainder of this paper, voucher "reception" and voucher "usage" are used interchangeably to refer to treatment.[3]

We then match the HUD data with the Master Address File Residential Auxiliary File (MAF-ARF). This dataset, maintained by the US Census Bureau linkage staff, provides the residential location of almost all Americans in each year between 2000-2019. The MAF-ARF is itself constructed by pooling data from multiple administrative sources: the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), the Indian Health Service database (IHS), the Selective Service System (SSS), the Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS), and the National Change of Address data from the US Postal Service.[4] As such, the MAF-ARF is the most comprehensive and accurate database on the annual residential location of individuals in the United States. Importantly, the MAF-ARF informs us on the living locations of individuals prior to the receipt of any housing voucher, and after the receipt of their last voucher.

To quantify the early life environmental exposure of children, we use data from the Census Household Composition key. This dataset links children 18 years and younger with their parents. We drop individuals whose residential location information is missing for more than five years between 2000 and 2019.[5] Our final dataset contains 2,231,260 unique women who received a housing choice voucher at any time between 2000 and 2018. Black voucher recipients account for 45% of voucher recipients in our sample.

Finally, we match each individual/year observation to a dataset of neighborhood characteristics resolved at the census block group level.[6] This newly developed product combines information from several sources to provide details on environmental and socio-economic measures of neighborhoods, such as yearly average air pollution (PM 2.5 and PM 10, from Dalhousie University's Atmospheric Composition Analysis Group), presence and releases from Toxic Release Inventory Sites (TRI, from the Environmental Protection Agency), distance and characteristics of municipal parks (from ParkServe), Annual Average Daily Traffic on roads (AADT, from the Department of Transportation Federal Highway Authority), flood risk (from FEMA and the First Street Foundation), schools test scores and diversity metrics (from GreatSchools), housing market

[2]Some information is available on voucher reception prior to voucher usage, but the relevant entries are missing for a large share of households. In ongoing work, we attempt to estimate the discrepancy between voucher receipt and usage.

[3]Within the set of records that consistently report the date of voucher reception, voucher reception and voucher usage generally track each other within a few weeks. Given that our residential outcomes are measured yearly, for the set of voucher users the effects of voucher reception and voucher usage are identical.

[4]Specifically, the Census data linkage staff extract from these sources an individual identifier called a PIK (for Protected Identification Key, which is a unique key used to link individuals across datasets in the Census data infrastructure) and a geokey called a MAFID (for Master Address File ID), which corresponds to a location on the Census Master Address file. Only individuals with a PIK appear in the MAF-ARF data, and each MAFID is associated with a State-County-Tract-Block FIPS code.

[5]A small share of individuals have missing residential information for more than 5 years. These results have yet to go through the disclosure process. We are currently investigating patterns of missingness in the MAF-ARF data.

[6]There are about 217,000 census block groups in the US, and each block group contains 1,200 individuals on average)

conditions (from Zillow) and a selected set of neighborhood-level metrics from the 2010 Decennial Census and pooled 2011-2015 ACS. Depending on availability and usage, these neighborhood-level measures are either static (measured only once, for instance in 2010 for Black population share) or dynamic (measured every year between 2000 and 2019).

Table 4.1 presents the consolidated dataset.[7] The average voucher recipient in our sample is 39 years old and received their first voucher in 2007. The average yearly moving rate between 2001 and 2019 is high, at 29%, and more than two thirds of these moves occur within the same county. Across all years, about 8% of individuals live in a census block group that contains a Toxic Release Inventory site. Individuals live in neighborhoods where the median household income is $41,890 (based on the pooled ACS 2011-2015 estimates). Race/ethnic groups are defined with a singular measure: "White" denotes White non-Hispanic, "Black" denotes Black non-Hispanic.

Table 4.2 presents summary statistics disaggregated by race, before receipt of the first housing voucher (indicated by *pre*) and after (*post*). Although our sample is composed of low-income individuals who all receive a housing voucher at some point, we can already see important disparities between race/ethnic groups. Before the receipt of their first voucher, White individuals live in neighborhoods with a median household income of $47,560 on average (against $36,070 for Black individuals), have access to greater park area within 2 kilometers of their neighborhood (134,200 $m^2$ vs 81,300 $m^2$), live further away from primary roads (2.2 km on average vs 1.1 km), are exposed to less road traffic coming from these roads (31,160 cars per day, vs 42,250 cars per day), and breathe cleaner air (10.45 vs 12.16 $\mu g/m^3$ of PM2.5). However, environmental inequalities are not one sided: White voucher recipients are more likely to live in a neighborhood that ever contained a toxic release inventory site (14.4% vs 11.9%) and are exposed to greater toxic releases from these sites (7,632 kg/year vs 5,459 kg/year, although this last difference is not significant due to the wide variation in toxic releases quantities). As we will see further below, such patterns of environmental inequalities are partly driven by the urban/rural divide: White individuals tend to live in more rural areas that are exposed to less air pollution but more toxic facilities than Black individuals, who live in denser city blocks.

Turning now to within-group comparisons before and after the receipt of the first voucher, it would appear from Table 4.2 that housing choice vouchers do not substantially change the probability of moving: for instance, for White voucher recipients the yearly moving average rate goes from 29% to 28%. Furthermore, looking at these averages alone, one might conclude that vouchers do not meaningfully improve most residential outcomes. For instance, prior to the receipt of their first voucher, Black individuals in our sample live in neighborhoods that are on average 1.1 kilometer away from a primary road, with a median household

[7] Flood risk measures, housing market conditions and schools outcomes are not yet displayed – US Census disclosure process ongoing

Table 4.1: Summary statistics

| Variable | Count | Mean | sd | Source | Measurement | Unit |
|---|---|---|---|---|---|---|
| Year | 41,390,000 | 2010 | 5.642 | n/a | Dynamic | n/a |
| Age | 41,390,000 | 39.06 | 16.99 | Decennial 2010 | Dynamic | n/a |
| Year relative child 1 | 20,940,000 | 3.982 | 7.043 | HH comp. key | Dynamic | n/a |
| Year relative child 2 | 16,350,000 | 1.997 | 7.285 | HH comp. key | Dynamic | n/a |
| Year relative child 3 | 11,220,000 | 0.326 | 7.42 | HH comp. key | Dynamic | n/a |
| Moving | 37,920,000 | 0.2947 | 0.4559 | MAF-ARF | Dynamic | 0/1 |
| Moving, different county | 37,920,000 | 0.08318 | 0.2762 | MAF-ARF | Dynamic | 0/1 |
| Assistance payment | 41,390,000 | 223.6 | 363.8 | HUD | Dynamic | USD |
| Male over 20 in hh | 15,980,000 | 0.1478 | 0.3549 | HUD | Static | 0/1 |
| Population, bg 2010 | 41,190,000 | 1598 | 978.7 | Decennial 2010 | Static | pers. |
| White pop. share, bg 2010 | 41,180,000 | 0.5735 | 0.3113 | Decennial 2010 | Static | Share |
| Black pop. share, bg 2010 | 41,180,000 | 0.2722 | 0.3119 | Decennial 2010 | Static | Share |
| Children below 9 share, bg 2010 | 41,180,000 | 0.147 | 0.04589 | Decennial 2010 | Static | Share |
| Median hh income, bg 2011 | 40,170,000 | 41890 | 20890 | Pooled ACS, 2011-2015 | Static | USD |
| Density, bg 2010 | 41,190,000 | 3179 | 7348 | Decennial 2010 | Static | pers / km2 |
| Nearest city at least 50K | 41,190,000 | 26.3 | 37.39 | Census Places | Static | km |
| Nearest primary road | 41,190,000 | 1527 | 2788 | Census TIGER/LINE, 2012 | Static | km |
| Nearest AADT | 41,190,000 | 39200 | 51510 | DoT FHA, 2011 | Static | cars/day |
| Nearest park area | 41,190,000 | 1744 | 4261 | ParkServe, 2019 | Static | km2 |
| Park area in buffer | 41,190,000 | 104900 | 561600 | ParkServe, 2019 | Static | m2 |
| PM 2.5 | 39,040,000 | 10.19 | 2.883 | ACAG | Dynamic | g/m3 |
| TRI, carcinogen presence | 41,190,000 | 0.05296 | 0.224 | EPA | Dynamic | 0/1 |
| TRI, count | 41,190,000 | 0.4785 | 3.463 | EPA | Dynamic | count |
| TRI, release on site | 41,190,000 | 5028 | 385100 | EPA | Dynamic | kg |
| TRI, presence | 41,190,000 | 0.08347 | 0.2766 | EPA | Dynamic | count |
| TRI, presence, ever | 41,190,000 | 0.1297 | 0.336 | EPA | Static | 0/1 |
| TRI, presence in tract, ever | 41,190,000 | 0.2995 | 0.4581 | EPA | Static | 0/1 |
| Admission date into program | 41,390,000 | 2007 | 5.02 | HUD | Static | year |
| Number of obs. per person | 41,390,000 | 18.55 | 2.303 | HUD+MARF | Static | count |
| Black | 41,390,000 | 0.4372 | 0.496 | Decennial 2010 | Static | 0/1 |
| White | 41,390,000 | 0.461 | 0.4985 | Decennial 2010 | Static | 0/1 |
| Hispanic | 41,390,000 | 0.06324 | 0.2434 | Decennial 2010 | Static | 0/1 |
| Asian | 41,390,000 | 0.01492 | 0.1212 | Decennial 2010 | Static | 0/1 |
| After admission | 41,390,000 | 0.6638 | 0.4724 | HUD | Dynamic | 0/1 |

Each observation is an individual-year. We focus on women household head, or on the household head's partner when the household head is a man. Race/ethnic groups are defined with a singular measure: "White" denotes White non-Hispanic, "Black" denotes Black non-Hispanic.

Table 4.2: Summary statistics by race, before and after receipt of voucher

| Variable | White pre | White post | Black pre | Black post | Hispanic pre | Hispanic post |
|---|---|---|---|---|---|---|
| Year | 2005 (4.179) | 2012 (4.748) | 2005 (4.155) | 2012 (4.815) | 2005 (4.175) | 2012 (4.773) |
| Age | 35.98 (18.47) | 45.28 (17.32) | 28.75 (14.57) | 39.21 (13.81) | 31.19 (15.13) | 41.76 (14.5) |
| Moving | 0.2883 (0.453) | 0.277 (0.4475) | 0.3232 (0.4677) | 0.3122 (0.4634) | 0.29 (0.4538) | 0.2597 (0.4385) |
| Moving, different county | 0.1037 (0.3048) | 0.08643 (0.281) | 0.08298 (0.2759) | 0.07445 (0.2625) | 0.08043 (0.272) | 0.06314 (0.2432) |
| Male over 20 in hh | 0.1599 (0.3665) | 0.1737 (0.3789) | 0.07208 (0.2586) | 0.1004 (0.3006) | 0.1659 (0.372) | 0.194 (0.3954) |
| Population, bg 2010 | 1572 (909) | 1612 (987.7) | 1468 (863.3) | 1611 (1021) | 1709 (980.9) | 1760 (1097) |
| White pop. share, bg 2010 | 0.7909 (0.1994) | 0.7783 (0.203) | 0.3358 (0.2697) | 0.3793 (0.272) | 0.5309 (0.2299) | 0.5235 (0.2331) |
| Black pop. share, bg 2010 | 0.07615 (0.1346) | 0.08225 (0.1389) | 0.5359 (0.3161) | 0.4831 (0.3116) | 0.1433 (0.1832) | 0.159 (0.1895) |
| Children below 9 share, bg 2010 | 0.1359 (0.04121) | 0.1384 (0.0446) | 0.1551 (0.04805) | 0.1562 (0.04594) | 0.1576 (0.04359) | 0.1573 (0.04285) |
| Median hh income, bg 2011 | 47560 (21710) | 44890 (20460) | 36070 (19180) | 37570 (19420) | 42180 (20800) | 42190 (20640) |
| Density, bg 2010 | 2038 (5330) | 2378 (5877) | 3244 (6569) | 3221 (6514) | 8927 (16350) | 8789 (15500) |
| Nearest city at least 50K | 35.61 (41.91) | 35.96 (42.69) | 18.18 (29.06) | 18.05 (28.85) | 15.8 (29.79) | 14.96 (28.88) |
| Nearest primary road | 2217 (3619) | 1959 (3409) | 1107 (2022) | 1069 (1817) | 1032 (1932) | 961.9 (1774) |
| Nearest AADT | 31160 (42400) | 31720 (43230) | 42250 (52990) | 43830 (54270) | 52000 (65440) | 50760 (63000) |
| Nearest park area | 2722 (5472) | 2371 (5228) | 1224 (3243) | 1124 (2868) | 844.3 (2504) | 762.6 (2356) |
| Park area in buffer | 134200 (767800) | 123400 (706400) | 81300 (337500) | 85990 (340900) | 82780 (418100) | 91150 (420200) |
| PM 2.5 | 10.45 (3.012) | 8.955 (2.608) | 12.16 (2.606) | 10.1 (2.355) | 11.82 (3.705) | 9.946 (2.956) |
| TRI, carcinogen presence | 0.06129 (0.2399) | 0.05852 (0.2347) | 0.05204 (0.2221) | 0.04944 (0.2168) | 0.03619 (0.1868) | 0.03108 (0.1735) |
| TRI, count | 0.5505 (3.492) | 0.4892 (3.173) | 0.5277 (4.044) | 0.4653 (3.601) | 0.329 (3.632) | 0.2633 (2.633) |
| TRI, release on site | 7632 (559600) | 5376 (524000) | 5459 (198100) | 3958 (164500) | 2473 (99660) | 1991 (152200) |
| TRI, presence | 0.09714 (0.2961) | 0.09278 (0.2901) | 0.08036 (0.2718) | 0.07715 (0.2668) | 0.05898 (0.2356) | 0.05251 (0.2231) |
| TRI, presence, ever | 0.1441 (0.3512) | 0.1466 (0.3537) | 0.1188 (0.3235) | 0.123 (0.3285) | 0.08989 (0.286) | 0.08817 (0.2835) |
| TRI, presence in tract, ever | 0.3405 (0.4739) | 0.3443 (0.4751) | 0.2702 (0.4441) | 0.2748 (0.4464) | 0.2125 (0.409) | 0.2069 (0.4051) |

Each observation is an individual-year. We focus on women household head, or on the household head's partner when the household head is a man. "bg" denotes census block group.

income of \$36,070, and which was home to at least one Toxic Inventory Site for 11.9% of individuals; after the receipt of a voucher, these statistics barely move to 1.1 km, \$37,570, and 12.3% respectively. As we are going to see in section 4.3 below, such pre/post comparisons are misleading.

## 4.3 Descriptive evidence on selection and treatment effects

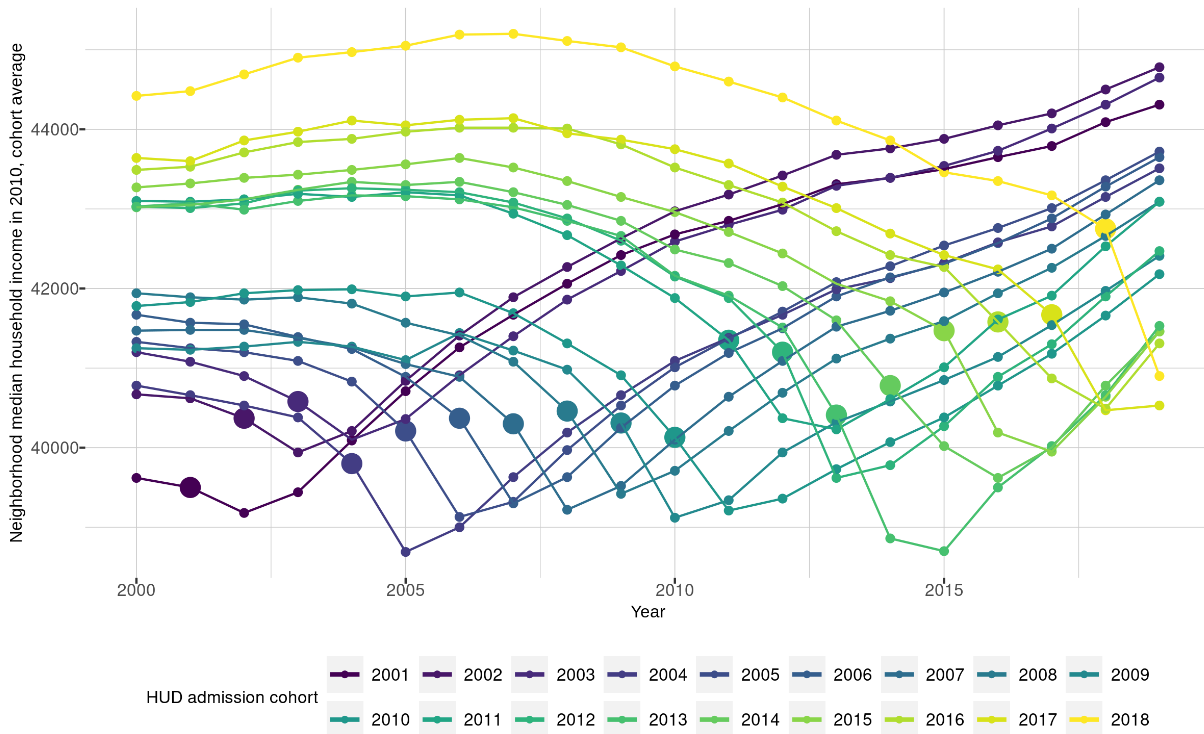
In this section, we leverage the entire residential history data for each individual housing choice voucher recipient to provide preliminary evidence on selection into treatment and inform our empirical strategy.

Figure 4.1 depicts the yearly average median income of the neighborhood where individuals live, in each year, broken down by cohort of housing voucher recipients (where a cohort is defined as the group of individuals who receive their first voucher in the same year). Two stylized facts emerge from these simple statistics.

### 4.3.1 Dynamic selection into treatment

First, there are striking dynamics prior to treatment: individuals tend to live in progressively *worse* neighborhoods until the receipt of their first housing choice voucher. The average neighborhood income of future voucher recipients starts to decline five to seven years *before* the receipt of the first voucher. This pattern is remarkably consistent between cohorts. Such decline of the outcome of interest prior to program par-

Figure 4.1: Preliminary evidence on dynamic selection into treatment and dynamic treatment effects




Each colored line represents the average median neighborhood income of housing voucher recipients for each voucher recipient cohort, where a cohort is defined as the group of individuals who receive their first voucher in the same year. The median neighborhood income is based on the pooled 2011-2015 ACS estimates and does not vary: any variation in the average median neighborhood income is entirely driven by the individual migrations of voucher recipients. The large dots on each line further shows the year in which the cohort starts receiving housing vouchers.

ticipation resembles the "Ashenfelter dip" initially discovered in the labor economics literature focused on estimating the impacts of training programs (Ashenfelter, 1978; Heckman and Smith, 1999). In our case, however, the decline starts several years prior to program participation instead of the several months decline typically observed in training programs. This pattern of selection into treatment resembles a long slide, rather than a dip.[8]

This decline in neighborhood quality prior to the receipt of a housing choice voucher is puzzling. Individuals are unlikely to voluntarily select into a poorer neighborhood "in anticipation" of receiving a voucher in the future. On the other hand, like job training programs, it might be that individuals apply to receive a voucher when their housing situation deteriorates. Unfortunately, the data do not allow us to test this hypothesis: the initial application date of voucher recipients is missing in most cases. However, a 2016 survey of Public Housing Authorities (PHA) by the National Low Income Coalition finds that out of 320 PHAs, 53%

[8]In ongoing work, we further separate results for individuals who stop receiving vouchers and those who keep receiving them for multiple years. We then estimate yearly differences in neighborhood quality at the individual level to highlight whether the observed average results are due to individuals moving at different times, or due to the same individuals moving multiple times.

of voucher waiting lists were closed to new applicants. For those who were able to apply, the median waiting time before receiving a voucher was 1.5 years, while a quarter of applicants had to wait more than three years (National Low Income Housing Coalition, 2016). Given these results, it might be that some individuals move to poorer neighborhoods even after being put on a waiting list to receive a housing voucher.[9] These patterns are consistent with the study of DeLuca et al. (2019) who fOund using interviews with low-income Black families that mobility was more *reactive* than voluntary, and often the results of evictions.

Selection into treatment also complicates the estimation of the effects of the program on the residential outcomes of voucher recipients. In the absence of randomization (such as in the Moving to Opportunity experiment, (Katz et al., 2001; Kling et al., 2007; Chetty et al., 2016; Ludwig et al., 2013)), one might hope to compare individuals who receive a voucher to individuals who are on a waiting list to receive a voucher but have not yet received it. However, the national HUD data files only provide records on individuals who ever *receive and use* a voucher, and even for these individuals, the waiting list entry date is missing in most cases. As an alternative, in section 4.4 below we use as a control group the individuals who *have not yet received* a housing choice voucher.

### 4.3.2 Effect dynamics

The second notable pattern from Figure 4.1 is the sharp dynamics: in the first two years following the initial receipt of a housing voucher, individuals in each cohort move to poorer neighborhoods than the census block groups they initially lived in. However, after two years, this pattern reverses and in each year after the third we observe increases in the average neighborhood income where individuals live. This pattern appears remarkably stable between years and cohort, with on average a $350 increase in median neighborhood income per year. The next section is devoted to the causal identification and estimation of these impacts.

## 4.4 Impact of receiving a voucher on residential outcomes

The combination of (i) the slide in cohort-neighborhood income prior to treatment, (ii) the apparent negative short-term impacts of the receipt of the voucher, and (iii) the seemingly constant increase in neighborhood income after two years is itself sufficient to rationalize why the naive comparison of residential outcomes pre and post voucher receipt in Table 4.2 would fail to find any detectable effect of the vouchers, even within race and ethnic group. Such patterns call for the estimation of dynamic treatment effects. This section

[9] While on waiting list, future housing choice voucher recipients might be offered to move to a public housing unit while they receive a housing choice voucher. If these individuals move to public housing located in neighborhoods that are poorer than where they used to reside, such migrations could contribute to the observed slide prior to treatment. We are currently using additional data to estimate how many individuals move to public housing prior to using a housing choice (i.e., “tenant-based”) voucher.

presents the empirical strategy we employ and the estimates of the impact of housing vouchers on residential outcomes.

### 4.4.1 Heterogeneity-robust difference-in-differences estimator

Different individuals receive vouchers in different years. To estimate treatment effects, it is therefore natural to employ a difference-in-differences estimator where we compare the evolution of the treated groups to the evolution of control groups. In an ideal (randomized) setting, such as the Moving to Opportunity experiment, we could simply compare the outcomes of individuals who receive vouchers to those of individuals eligible to receive vouchers but who did not receive any. In our context, we only observe individuals who actually receive and use a voucher at some point between 2000 and 2018. A first possibility could be to use additional data and to match individuals who receive vouchers to a different sample of individuals who do not receive any housing voucher (for instance using microdata from American Community Surveys), and then use these individuals as controls by matching on observables or propensity score. However, such strategy is unlikely to be fruitful for at least two reasons: (i) the previous section highlights patterns that are strongly suggestive of dynamic selection into treatment, and matching on observables is a priori unlikely to deliver such specific patterns, and (ii) housing voucher recipients are different from the US population in systematic ways (in particular, Black and Extremely Low Income individuals are disproportionately-represented), with some relevant characteristics likely to be unobservable.

The alternative we employ in this section is to use as control units the individuals who *have not yet* received a housing voucher. Given that eligibility criteria to receive a housing voucher have not fundamentally changed over the past two decades, this approach is more likely to generate a control group that is similar to the treatment group. On the other hand, our estimates are only valid for the subset of voucher recipients who successfully use their vouchers – individuals who are offered a voucher but do not use it are not represented in our dataset.

Within the *not-yet-treated* individuals, one additional important choice that we need to make concerns which specific not-yet-treated individuals should be used as control units for individuals that receive a housing choice voucher earlier. Based on the previous section, this choice implies an important trade-off between (i) the quality of the matching on the dynamics of the selection into treatment on the one hand, and (ii) the number of post-period treatment effects that we are able to estimate on the other hand. To clarify, consider the cohort of individuals who received their first housing choice voucher in 2005. To estimate the effects on individuals in this cohort, one possibility is to form a control group composed of only individuals who received their housing choice voucher in 2018. This would make it possible to estimate the dynamics of treatment effects over a 13-year period ($2005-2018$). However, due to the dynamics of selection into treatment noted above, the parallel trend assumption for the untreated potential outcomes is likely to be violated: for the cohort of individuals treated in 2018, residential income starts to decline about 9 years prior, in 2010, and is relatively flat (with even a slight increase) prior to 2005. The other "extreme" choice would be to use as controls only the individuals that receive treatment in 2006. In this case, the parallel-trend assumption would be likely to hold prior to 2005, but only one year of post-treatment effect could be estimated (for 2005 only, since in 2006 both groups will be treated).

Our preferred specification, presented below, uses as control units *all* the individuals who have not yet received a voucher at the time of their treatment. This allows us to estimate a large number of effects post-treatment. We use the common test for rejection of the parallel trend assumption by estimating the set of pre-treatment coefficients and test whether they are systematically different from zero. Because individuals from different race/ethnic groups tend to live in systematically different neighborhoods and might be subject to different types of discrimination, we estimate treatment effects separately for Black, White, and Hispanic voucher recipients. Finally, a recent econometric literature highlights important pitfalls with the use of the two-way fixed effect and standard event study estimators when treatment is staggered (de Chaisemartin and D'Haultfoeuille, 2020; Sun and Abraham, 2021). In particular, when treatment effects are heterogeneous between cohorts, the commonly used event study estimators generically do not recover the Average Treatment Effect on the Treated (ATT). Applied work further found that these issues are not merely theoretical, but that standard event study estimators often provide results that have the "wrong" sign (Baker et al., 2021; **?**). We follow the approach of Callaway and Sant'Anna (2020) and first estimate treatment effects independently for each cohort before re-aggregating them to provide heterogeneity-robust event study estimates. Formally, we are interested in the average treatment effect on the treated for each group $g$, time period $t$, and race/ethnic group $x$:

$$ATT(g,t,x) = \mathbb{E}(Y_t(g,x) - Y_t(0,x)|G_g = 1, X = x) \tag{4.1}$$

where $Y_t(g,x)$ denotes the potential outcome of units in group $x$ at time $t$, *if* they were to be part of cohort $g$ (and hence become treated at time $t = g$). $Y_t(0,x)$ denotes the potential outcome of these same units had they not received a treatment. For each unit we can only observe the *realized* outcome: to estimate the ATT, the expectation over potential outcomes is replaced with sample averages (where we choose the control group to be made of units that have not yet received treatment at time $t$). Finally, we further follow Callaway and Sant'Anna (2020) and aggregate the group-time ATTs to provide estimates based on the length of exposure to treatment:

$$\theta(e,x) = \sum_g \mathbf{1}\{g+e \le T\} \cdot P(G=g \ \cap \ X=x|G+e\le T\} \cdot ATT(g, g+e, x) \tag{4.2}$$

where e is the exposure time (e = t - g), g is an index for the cohort of first-year voucher recipients, G is the time period that a unit is first treated, and T is the last time period for which ATTs are identified.

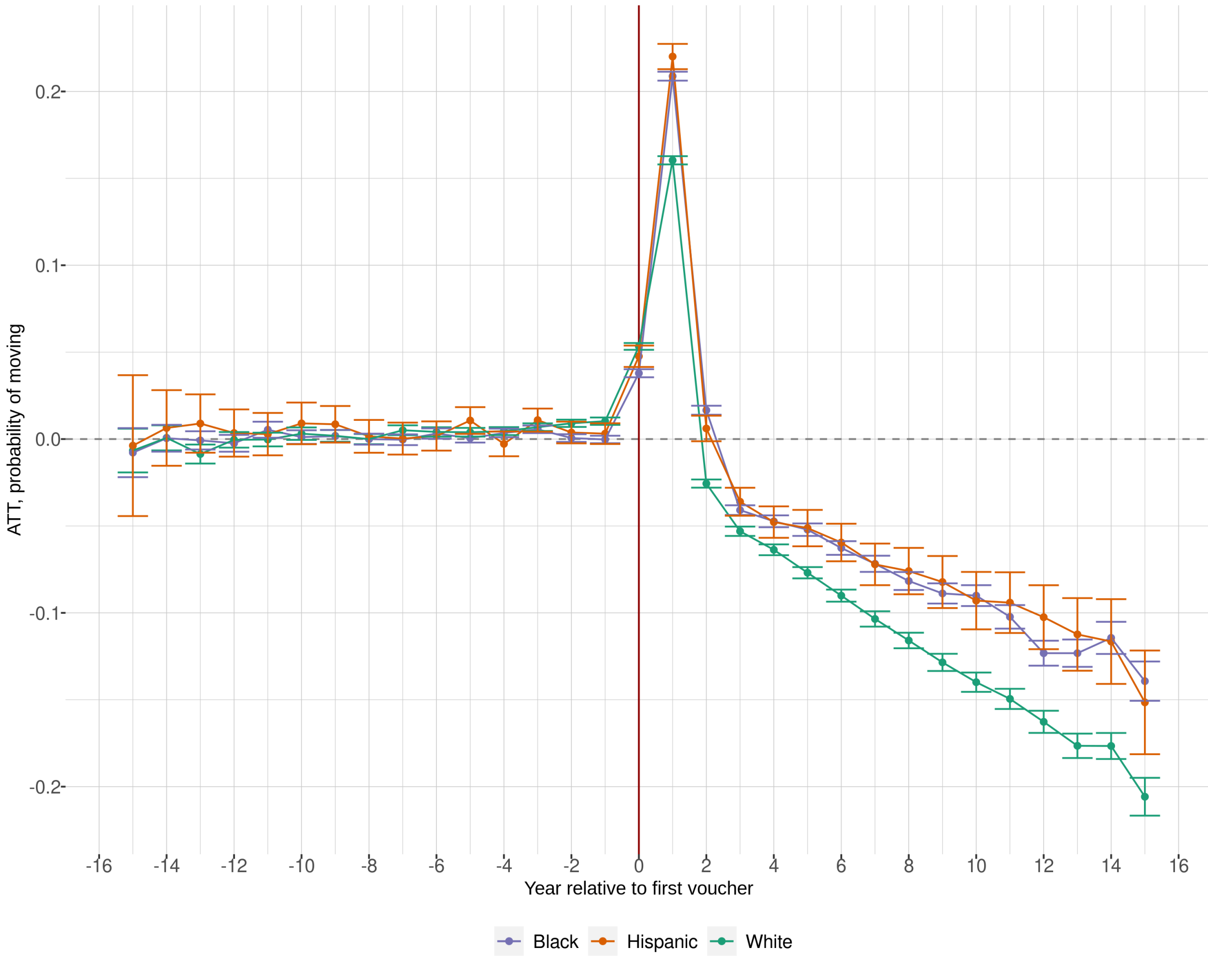
### 4.4.2 Results from the event studies

Figure 4.2 shows the estimated impacts of the receipt of the initial housing voucher on the probability of moving per race and ethnic group, using observations from 2001 to 2018.[10] For all three race/ethnic groups, the estimated ATTs are not significantly different from zero prior to treatment and we fail to reject the parallel trend assumption. Immediately in the year of the receipt of a voucher, we estimate the probability of moving increases by 5 percentage points for the three groups we consider (relative to 16 years prior to treatment, which serves as the baseline year). One year after receiving a voucher, the probability of moving increases by 20 percentage points for Black and Hispanic voucher recipients, and by 16 percentage points for White voucher recipients. These preliminary checks provide convincing evidence that vouchers are actually used to move to different neighborhoods. Most of these moves are within-county: Figure J.2 panel 1 shows that the shape of the effects is almost identical for between-county moves, but that the magnitude is much smaller (with a peak of 3 percentage points increase for Black and Hispanic families). Interestingly, the treatment effects become null or negative already in year 3, and monotonously decline until the end of the estimation window (15 years post treatment). Recall from section 4.2 that the yearly migration rate prior to receiving a housing choice voucher is extremely high, around 30% for the three groups we consider. Receiving a voucher temporarily increases this migration rate (immediately after the receipt of the first voucher), and causes individuals to migrate *less* in the following years. These results are consistent with increased housing stability – vouchers enable households to change neighborhoods in the short run, and allow them to migrate less in the medium and long run (relative to the pre-voucher baseline period and the families who haven't yet received their vouchers). However, the medium and long run dynamics are not characterized by complete immobility – as the next results show, voucher recipients continue to migrate after the receipt of their voucher, but these migrations are less frequent.

Figure 4.3 begins the presentation of the impacts on the residential outcomes of voucher recipients by showing the ATTs on average neighborhood income. Recall that income is measured from the pooled ACS 2011-2015 and is *fixed throughout our observation window*: all variations in this measure therefore come from the spatial sorting of voucher recipients. The pre-treatment coefficients are slightly negative up to four years

[10] Given that 2000 is the first year of our sample, we do not know which individuals moved in that year

Figure 4.2: Average treatment effects estimates of housing choice vouchers on the probability of moving

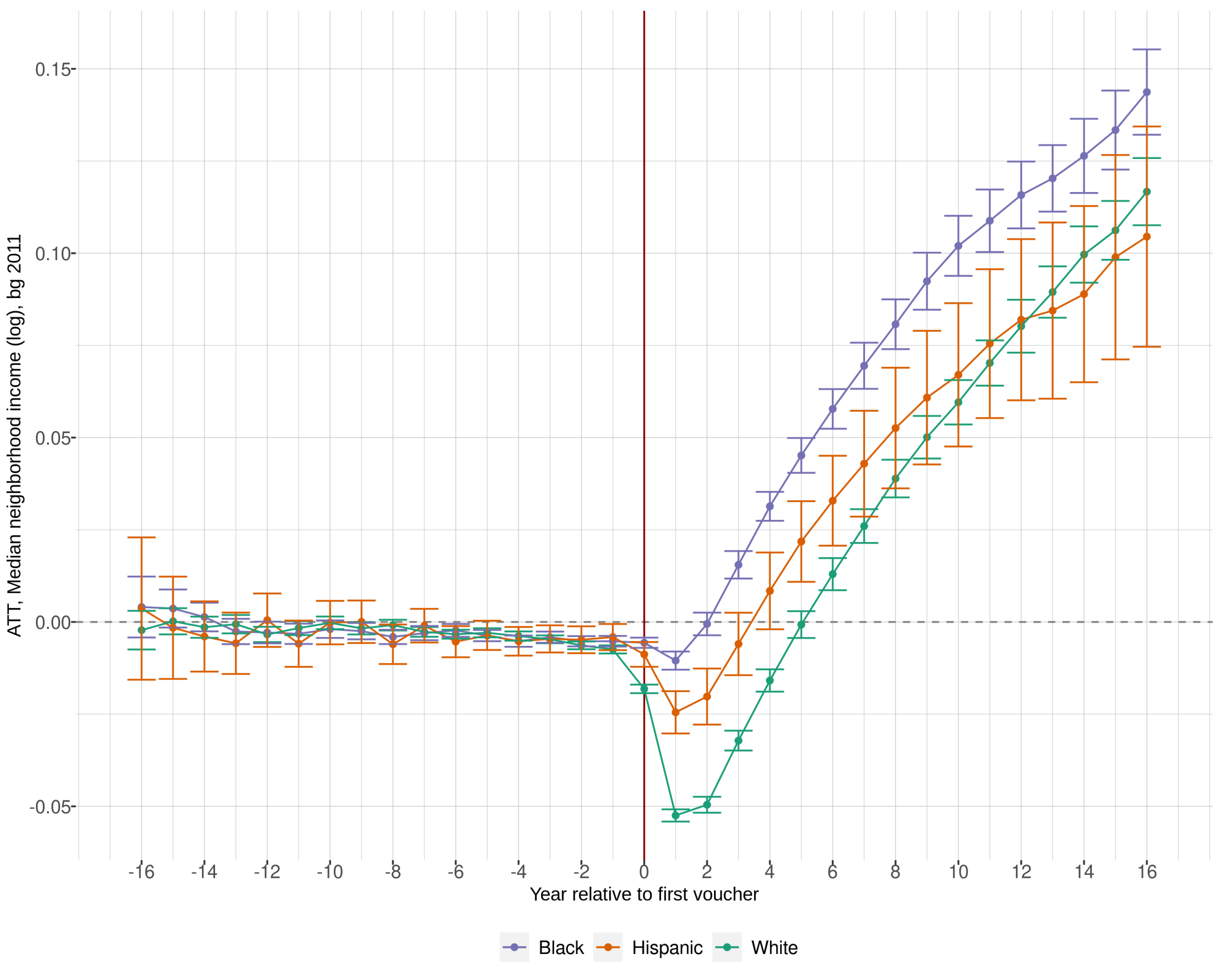



The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap.

prior to treatment, which reveals that using all not-yet-treated individuals as controls might induce some comparisons between individuals who are at different points of the pre-treatment dynamic selection process. These pre-trends are however extremely small, and section 4.4.4 below shows that they disappear when running event studies by state, providing further support for the causal interpretation of these estimates.

Immediately after the receipt of their first voucher, treatment effects become negative. They remain negative for 5 years for White voucher recipients, while the effect lasts 3 and 2 years for Hispanic and Black families, respectively. The negative effect is also stronger for White families: these households move to neighborhoods that are on average 5% *poorer* than the neighborhood they started from. However, after 1 year post treatment, treatment effects increase for all three groups, at a relatively constant rate of 1.25% per year. At the end of our estimation window, individuals who experienced 16 periods post treatment (those treated in 2001) live in neighborhoods that are between 10 and 14% wealthier than the neighborhoods they lived in 2000, compared to the control group composed of individuals who received their voucher in 2018. Because Black voucher recipients start from relatively poorer neighborhoods than White and Hispanic voucher recipients, they experience a less negative treatment effect in the short-run, and therefore experience greater relative increases in neighborhood income in the long run. Overall, these results provide conclusions that are extremely similar to the patterns in the raw data presented above.

Figure 4.3: Average treatment effects estimates of housing choice vouchers on neighborhood income




The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap. Income is measured as the median income from the pooled ACS 2011-2015.

Combining the results from Figures 4.2 and 4.3 reveals that vouchers increase mobility in the short run, but these moves are directed towards poorer neighborhoods on average. After the first couple of years vouchers cause families to migrate *less*, while simultaneously enabling them to move to higher income neighborhoods.

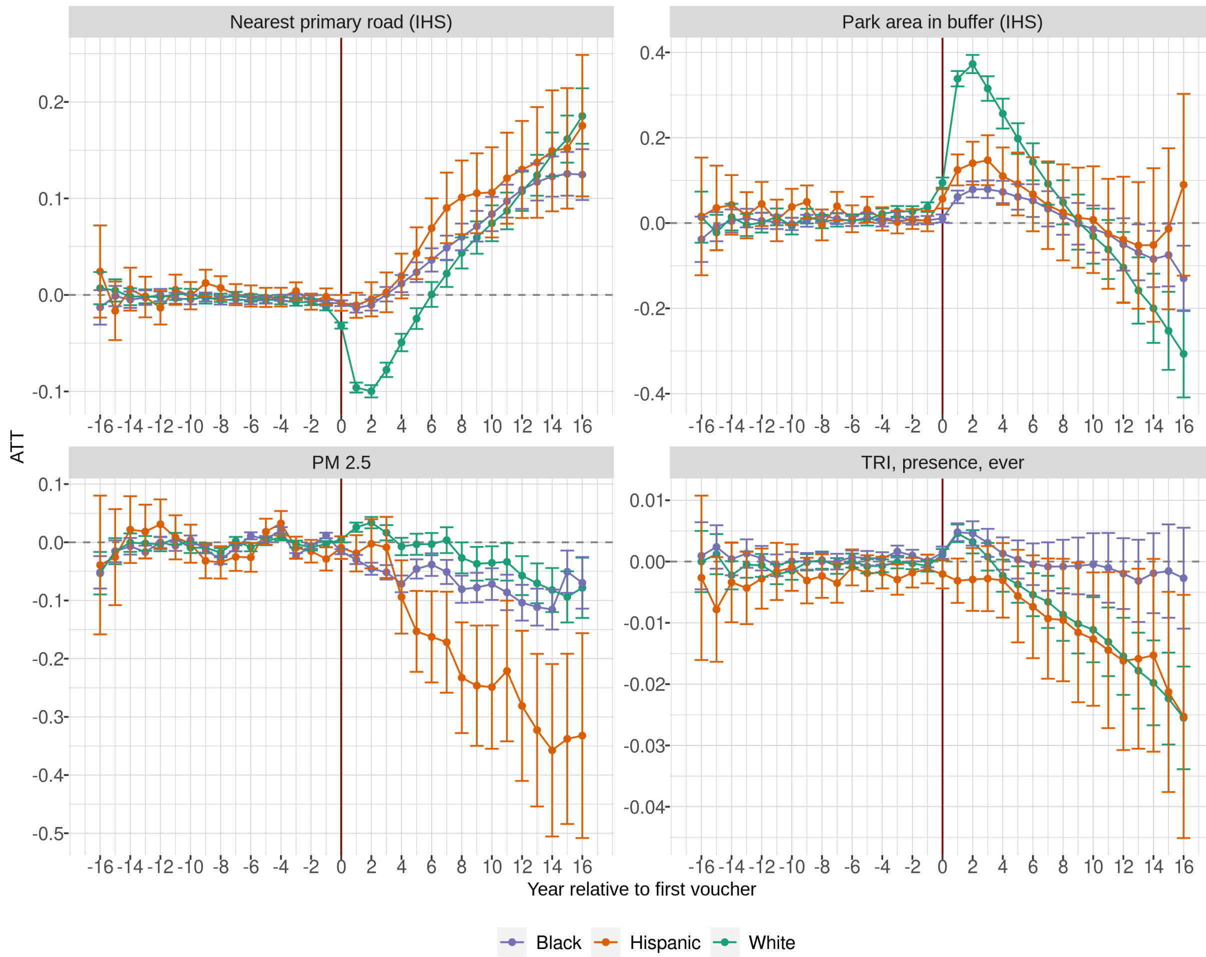
Why are treatment effects negative in the short run, and positive after a few years? Two mechanisms are consistent with these dynamics. First, using vouchers is hard: both parties (the tenant and the landlords) need to agree in order for the voucher to be used, and previous work found evidence of multiple barriers that prevent voucher holders from using vouchers to move to better neighborhoods (Kleit et al., 2015; Phillips, 2017; Bergman et al., 2019; Ellen, 2020). Voucher users might progressively learn about the program and how to find suitable housing in better neighborhoods after their first year in the program. A second possibility is that families are "forced" to use their vouchers in poorer neighborhoods because of the vouchers' expiration provisions: while the precise date of validity of vouchers may vary between Public Housing Agencies, in some areas housing choice vouchers are only valid for 60 days – when they expire, the voucher cannot be used anymore and households have to re-apply for the PHA waiting list. This provision likely compels voucher recipients to use their voucher as quickly as possible in order to keep it, and for many families this means accepting to move to a poorer neighborhoods where landlords are more than happy to rent to voucher holders as this guarantees a steady flow of rental income (Rosen, 2020). In ongoing work, we attempt to disentangle these mechanisms.

### 4.4.3 Effects on broader neighborhood characteristics

We now investigate the impacts of housing choice vouchers on environmental outcomes. Figure 4.4 presents event study estimates for the distance to the nearest primary road, access to municipal parks (measured as the park area within a 2km radius around the neighborhood centroid), exposure to PM2.5, and the presence of a toxic release inventory (TRI) site in the neighborhood (at any point between 2001 and 2019).

First, regarding the exposure to PM2.5 (the only variable measured yearly among the four presented here), post-treatment effects are negative and statistically significant, with dynamics revealing gradual decreases in exposure to PM2.5: -0.1$\mu g/m^3$ for Black and White families and -0.35 $\mu g/m^3$ for Hispanic families after 15 years.

Figure 4.4: Average treatment effects estimates of housing choice vouchers on selected environmental outcomes




The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap. IHS denote the Inverse Hyperbolic Spline transformation, used for fat-tailed outcomes.

Turning to our static measures of environmental outcomes, we find that vouchers cause all three race/ethnic groups to move further away from primary roads and to initially gain access to greater municipal park area. The effects display strong dynamics, especially for White voucher recipients, who initially move closer to primary roads. Effects on the presence of a TRI site in the neighborhood are not statistically significant for Black and Hispanic households, while they are negative and precisely estimated for White households, with a decline of 2.5 percentage point after 15 years (a substantial 18% decline relative to their pre-treatment average). Overall, housing vouchers appear to slightly improve environmental outcomes for all three groups, with especially large improvements for White voucher recipients.

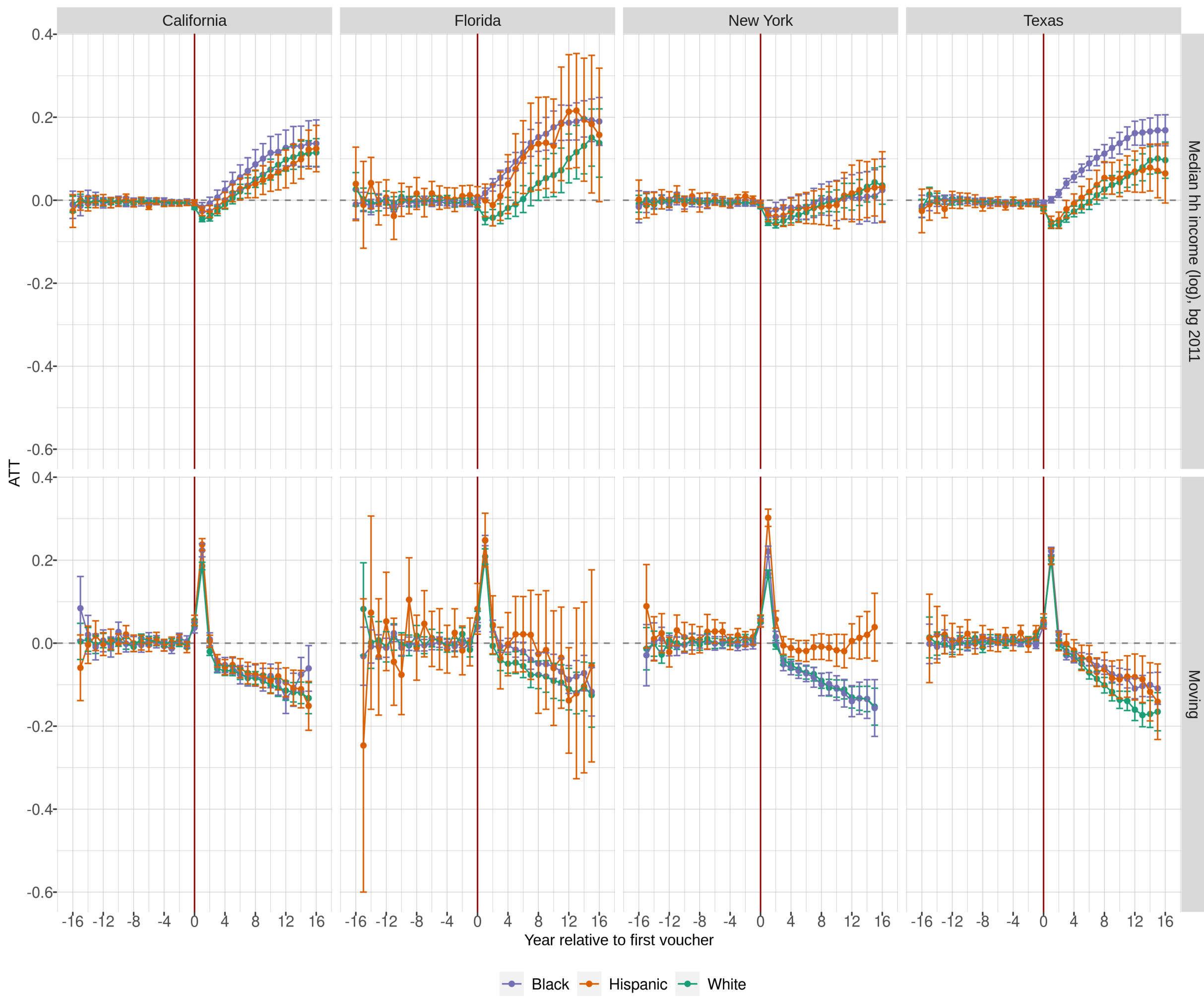
Figure J.1 in the Appendix provides some insights into the origins of these effects by displaying the event study estimates for additional neighborhood outcomes. The first and third panels show that housing vouchers tend to reduce segregation: they cause Black individuals to move to neighborhoods with a lower share of Black residents and greater share of White residents. White voucher recipients tend to initially move to more diverse neighborhoods that are also more urban, at least in the first 10 years after the receipt of the voucher. Given that urban neighborhoods are less likely to have a TRI site, this pattern partially explains the results presented above. However, the sign of the average treatment effect reverses after 10 years. The next section presents estimates disaggregated by state to gain further insights into these results.

### 4.4.4 State-specific ATTs

The previous event studies presented estimates of average treatment effects over the entire United States. Although these average effects are within race/ethnic groups, interpreting the impacts on several residential outcomes is tricky; neighborhood characteristics can vary widely by state. For instance, PM2.5 concentrations in California were about 30% higher than the national average in 2019, while they are slightly below the national average in Florida. These between-states differences in environmental outcomes can then become conflated with between-states differences in housing choice vouchers usage rates: if certain states have more voucher recipients in certain years, then the dynamic treatment effects estimated over all states will conflate dynamics of the spatial sorting process with changes in compositions in the relative share of each state to the national average. To disentangle these effects, in this section we re-estimate ATTs independently for each state-race/ethnic group. We focus on the four states with the largest number of voucher recipients: California, Texas, New York and Florida.

Figure 4.5 presents the impact of receiving the initial housing choice voucher on the probability of moving and the average median neighborhood income by state. As before, each panel represents dynamic ATTs that were estimated on a different subsample (here race/ethnic group within each state). We note that the pattern observed above in Figures 4.2 and 4.3 over the national race/ethnic groups sub-samples continue to hold within each state: the short-term treatment effect on the probability of moving is large (at about 20% for individuals in each group), and becomes negative after the second year. Similarly, the short-term impacts on neighborhood income are initially negative, but monotonously increase until the end of our estimation window. Interestingly, the long term impacts are positive for all race and ethnic groups in California, Florida and Texas, while the impacts are small and not significantly different from zero in New York. In ongoing work, we investigate why vouchers appear to perform significantly worse on neighborhood outcomes in New York compared to other states.

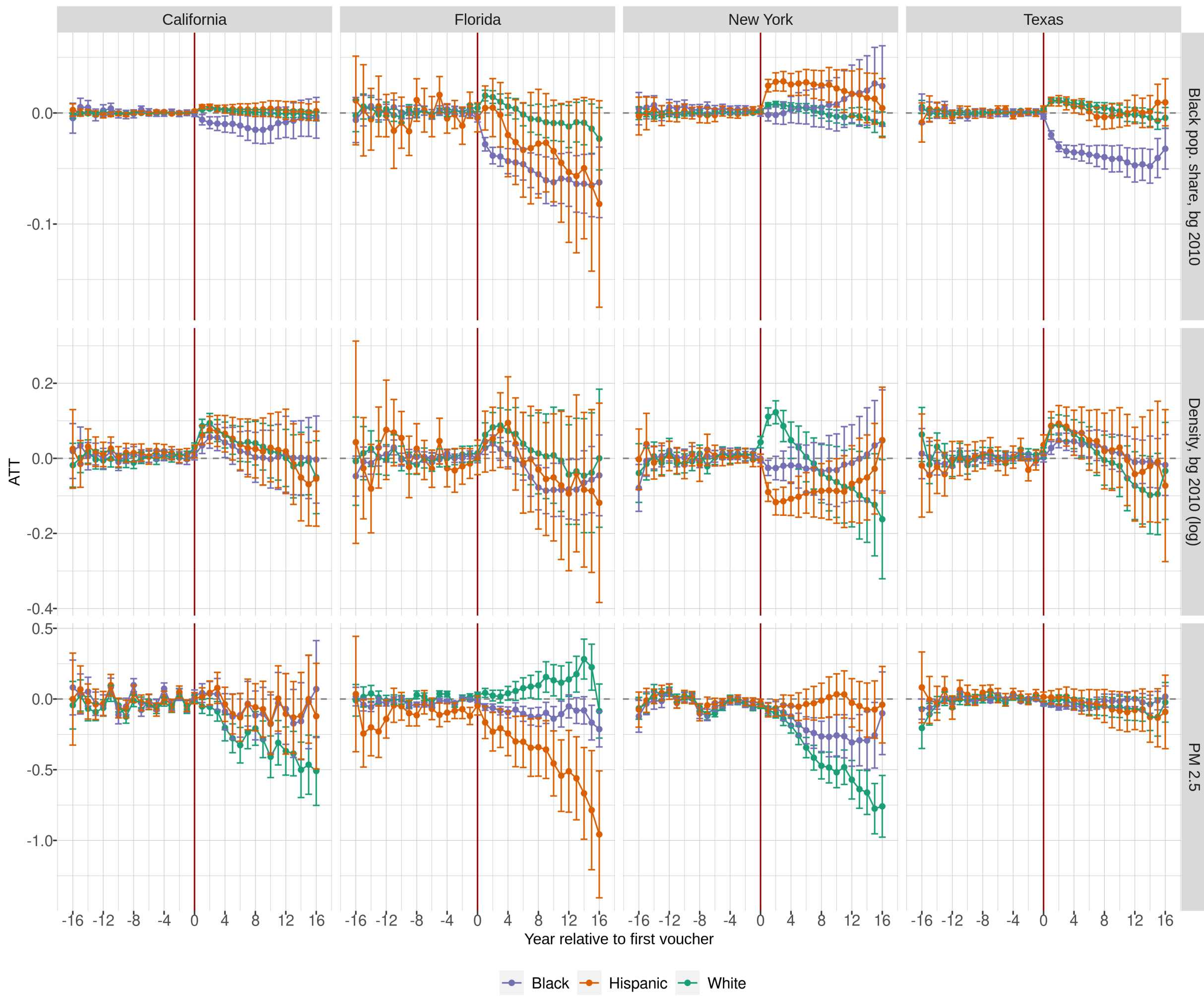
Figure 4.5: Average treatment effects estimates of housing choice vouchers on the probability of moving and neighborhood income, by state




The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap.

Figure 4.6 presents state-level event study estimates for three additional neighborhood outcomes: Black population share, population density, and PM2.5 concentration. Black voucher recipients move to neighborhoods with a smaller share of Black residents in Florida and Texas, while there does not appear to be any effect in California and New York (first row). White voucher recipients appear to disproportionately gain access to cleaner air in California and New York following the receipt of the housing choice vouchers (with effects of -0.5 $\mu g/m^3$ and -0.75 $\mu g/m^3$ after 15 years, respectively), while vouchers appear to increase the pollution exposure of White families in Florida.

Figure 4.6: Average treatment effects estimates of housing choice vouchers on selected environmental outcomes, by state




The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap.

Overall, the previous analysis highlights that effects of vouchers on residential outcomes exhibit strong dynamics and vary between states. While vouchers cause Black individuals to move to wealthier and more White neighborhoods, vouchers in New York do not appear to increase integration.

## 4.5 Discussion

This paper revisits the impacts of housing choice vouchers on the residential outcomes of beneficiaries by consolidating novel records comprising the universe of voucher users linked with residential address files and a novel dataset of neighborhood outcomes. We find evidence of dynamic selection into treatment: future voucher users move to progressively poorer neighborhoods prior to the receipt of their first vouchers. This pattern is stable between cohorts of families receiving vouchers at different times, and reveals that pre-post comparisons of the residential outcomes of tenants are likely to underestimate the effects of housing choice vouchers.

Using newly developed difference-in-difference models, we find that the impacts of vouchers on residential outcomes are strongly dynamic: voucher users initially move to poorer neighborhoods, while they progressively move to wealthier and less polluted neighborhoods after a few years. While effects are consistent between race and ethnic groups, they are differentiated between states, with voucher users in New York state having no detectable effects on several measures of neighborhood quality.

The short-run negative impacts of vouchers on neighborhood quality are surprising. These effects could be rationalized by a learning mechanism, or could be due to the "use-it-or-lose-it" provision attached to many vouchers, which might force families to accept sub-optimal housing options in order to prevent their housing voucher from expiring. In ongoing work, we investigate both mechanisms and leverage variation on the receipt of multiple vouchers to disentangle the treatment effect dynamics between cumulative effects and length-of-exposure effects.

## 4.6 References


Alexander, D. and J. Currie (2017). Is It Who You Are or Where You Live? Residential Segregation and Racial Gaps in Childhood.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economic and Statistics 60*(1), 47–57.

Baker, A., D. F. Larcker, and C. C. Y. Wang (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? *SSRN Electronic Journal* (March).

Bergman, P., R. Chetty, S. Deluca, N. Hendren, L. F. Katz, and C. Palmer (2019). Creating Moves To Opportunity: Experimental Evidence on Barriers to Neighborhood Choice. *NBER Working Paper*.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics, forthcoming*, 1–31.

Center on Budget and Policy Priorities (2020). The Housing Choice Voucher Program. Technical report, CBPP.

Chetty, R. and N. Hendren (2018). The Impacts of Neighborhoods on Intergenerational Mobility I:Childhood Exposure Effects. *Quarterly Journal of Economics 133*(3), 1107–1162.

Chetty, R., N. Hendren, and L. F. Katz (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review 106*(4), 855–902.

Chyn, E. and L. F. Katz (2021). Neighborhoods matter: assessing the evidence for place effects. *NBER Working Paper*.

Collinson, R. and P. Ganong (2018). How do changes in housing voucher design affect rent and neighborhood quality? *American Economic Journal: Economic Policy 10*(2), 62–89.

Colmer, J. and J. Voorheis (2020). The grandkids aren't alright: the intergenerational effects of prenatal pollution exposure. *Working paper*.

de Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-Way Fixed Effects Estimators With Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

DeLuca, S., P. M. Garboden, and P. Rosenblatt (2013). Segregating Shelter: How Housing Policies Shape the Residential Locations of Low-Income Minority Families. *Annals of the American Academy of Political and Social Science 647*(1), 268–299.

DeLuca, S. and P. Rosenblatt (2017). Walking Away From The Wire: Housing Mobility and Neighborhood Opportunity in Baltimore. *Housing Policy Debate 27*(4), 519–546.

DeLuca, S., H. Wood, and P. Rosenblatt (2019). Why Poor Families Move (And Where They Go): Reactive Mobility and Residential Decisions. *City and Community 18*(2), 556–593.

Ellen, I. G. (2020). What do we know about housing choice vouchers? *Regional Science and Urban Economics 80*, 103380.

Heckman, J. J. and J. A. Smith (1999). The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies. *The Economic Journal 109*(457), 313–348.

Holloway, A. M. (2014). From the City to the Suburbs: Characteristics of Suburban Neighborhoods Where Chicago Housing Choice Voucher Households Relocated. *Urban Studies Research 2014*, 1–14.

Isen, A., M. Rossin-Slater, and W. R. Walker (2017). Every breath you take - every dollar you'll make: the long term consequences of the Clean Air Act of 1970. *Journal of Political Economy 125*(3).

Katz, L. F., J. R. Kling, and J. B. Liebman (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics 116*(2), 607–654.

Kleit, R. G., S. Kang, and C. P. Scally (2015). Why Do Housing Mobility Programs Fail in Moving Households to Better Neighborhoods? *Housing Policy Debate 26*(1), 188–209.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling, and L. Sanbonmatsu (2013). Long-term neighborhood effects on low-income families: Evidence from moving to opportunity. *American Economic Review 103*(3), 226–231.

National Low Income Housing Coalition (2016). The long wait for a home. Technical report.

Pendall, R. (2000). Why voucher and certificate users live in distressed neighborhoods. *Housing Policy Debate 11*(4), 881–910.

Phillips, D. C. (2017). Landlords avoid tenants who pay with vouchers. *Economics Letters 151*, 48–52.

Rosen, E. (2020). *The Voucher Promise: "Section 8" and the Fate of an American Neighborhood.* Princeton University Press.

Sun, L. and S. Abraham (2021). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *Journal of Econometrics 225*(2), 175–199.

Tighe, J. R., M. E. Hatch, and J. Mead (2017). Source of Income Discrimination and Fair Housing Policy. *Journal of Planning Literature 32*(1), 3–15.

Tobias, M. and J. Botts (2019). Thousand of Californian Renters with Section 8 Vouchers Can"t Use Them: What Lawmakers are Doing About it.

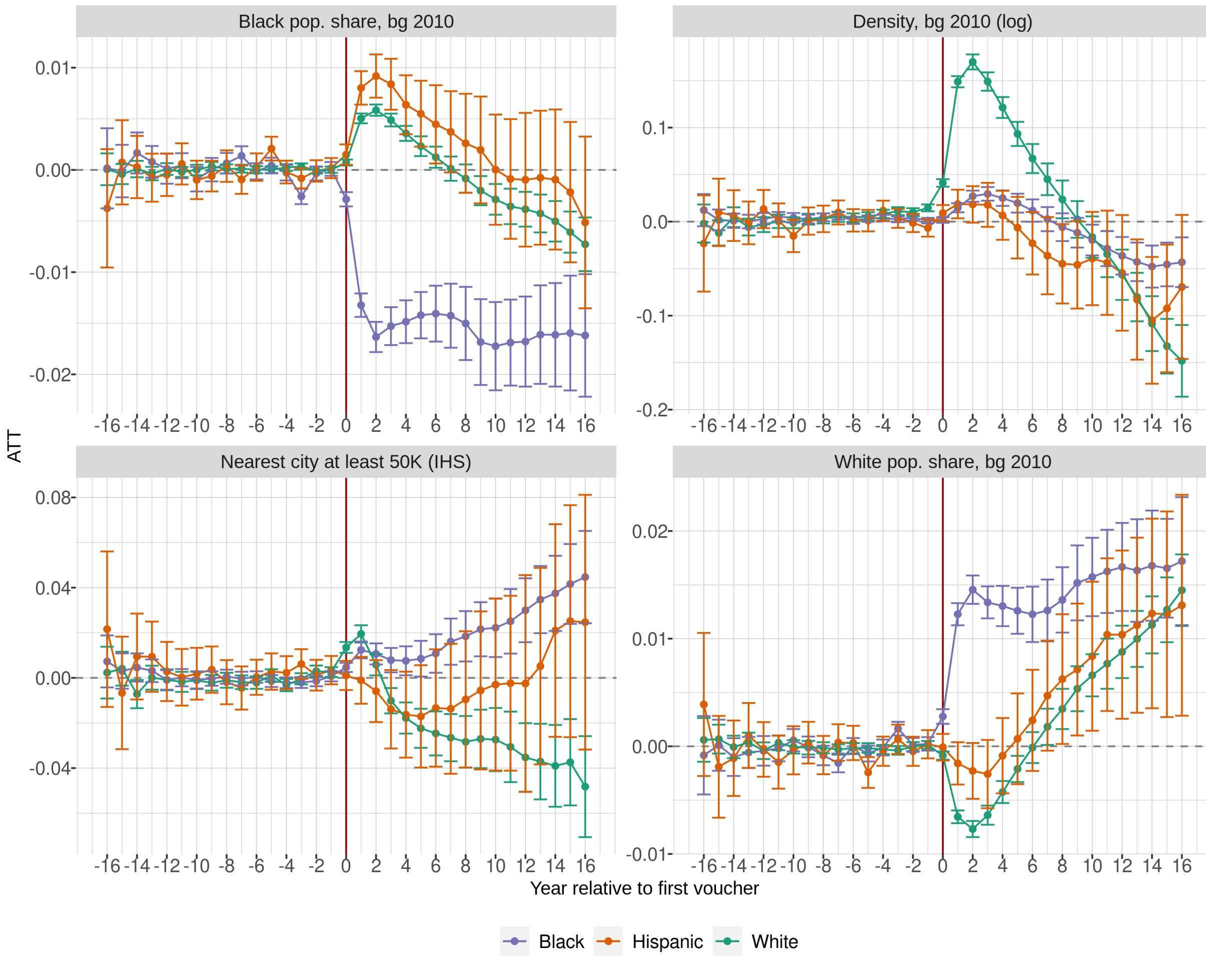
Versoulis, A. (2020). 'A Mask for Racial Discrimination.' How Housing Voucher Programs Can Hurt the Low-Income Families They're Designed to Help.

# Appendices

# Appendix J

# Additional results on the impacts of housing choice vouchers

Figure J.1: Average treatment effects estimates of housing choice vouchers on selected outcomes




The cohort-specific ATTs are aggregated by length of exposure. Error bars depict 95% confidence intervals based on a multiplier bootstrap.

Figure J.2: Average treatment effects estimates of housing choice vouchers on additional outcomes