Title
Essays in Urban Economics and Spatial Econometrics

Author
Aljutaili, Dhari S. R. S. E.

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
IRVINE

Essays in Urban Economics and Spatial Econometrics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Dhari S. R. S. E. Aljutaili

Dissertation Committee:
Professor David Brownstone, Chair
Professor Jan K. Brueckner
Associate Professor Ivan Jeliazkov
Kevin D. Roth, Ph.D.

2018

DEDICATION

To my mother, Dhiaa,

and my wife, Eman,

for their love, support, and sacrifice.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

This research would not have been possible without the support of my dissertation committee members. Meeting Professor David Brownstone, my committee chair, during the recruitment visit in March of 2013 was one of the reasons I chose to attend UCI. I am always awed by his depth of knowledge and experience in economics and other fields. I am equally appreciative of his guidance and calming effect in every challenge I was confronted with in my research. I am honored to have been David's last student before he retired and wish him a happy and well-deserved retirement.

Professor Jan Brueckner's scholarly contributions and neat lectures solidified my interest in urban economics. His capacity to take genuine interest in my research, reading drafts meticulously and providing constructive feedback, is something I hope to emulate with my students. I greatly appreciate his spreading the word about my work, most times unbeknownst to me. Hearing someone in the department or at a conference say *"Jan told me about your paper"* was always a pleasant surprise.

I am a Bayesian because of Professor Ivan Jeliazkov, who showed me how Bayes works and why it makes more sense. It was a privilege to know the secret door knock to which Ivan always answered and generously took the time to discuss my work in his office. I especially appreciate the hours of coding I was spared multiple times because he knew of a shortcut function that accomplished what I needed.

Professor Kevin Roth has been integral to my transition into becoming an economist. I owe the coherence and economic grounding in my way of thinking and writing to his challenging me with tough, fundamental questions in the kindest, most constructive ways possible. I appreciate Kevin staying on my committee after he left UCI and wish him the best.

I recognize funding provided by David Brownstone, the Department of Economics, and the School of Social Science to support my research. I thank my professors and peers in the department for making it a place where knowledge was not only gained but also enjoyed. I am especially thankful to my study group mates, Kelsey Heider and Jessica Monnet, for making the first two years survivable. I will dearly miss my weekly lunch with Jessica.

I am indebted to my home country, Kuwait, for the countless opportunities I was given in life, including my education from the first day of kindergarten to the last day of my Ph.D. While the causes and consequences of the resource curse remain a subject of debate among economists, it surely has been nothing but a blessing to this son of a middle-class teacher and single mother. I look forward to returning home and giving back.

I owe a debt of gratitude to my family for their positive vibes from half the world away: Mama Latifah's prayers, my wife's good-luck wishes, and my mother's pride. I thank my friends Dalal Alfares and Abdullah Husain for the travels, group chats, and shared moments of our graduate school endeavors. Finally, I am thankful to Loomi, our sweet dog, for the joy, comfort, and companionship that eased the stresses of the last two years.

On to the next adventure...

CURRICULUM VITAE

Dhari S. Aljutaili

EDUCATION

University of California, Irvine

Doctor of Philosophy in Economics 2018

Master of Arts in Economics 2015

Oregon State University

Master of Science in Mechanical Engineering 2007

Bachelor of Science in Mechanical Engineering with a Minor in Entrepreneurship 2005

RESEARCH FIELDS

Urban and transportation economics.

Applied spatial, discrete choice, and Bayesian econometrics.

RESEARCH

Working Papers

"The Relationship Between Neighborhood Design and Social Capital as Measured by Carpooling," (under review).

"Modeling Panel Count Data with Dynamics and Spatial Correlation."

"Peer Effects and Spatial Correlation in Solar Panel Adoption."

Research Experience

Research assistant to Profs. Linda Cohen & Kevin Roth, Dept. of Economics, UCI. 2016

CONFERENCE & SEMINAR PRESENTATIONS

Annual Meetings, Urban Economics Association, Minneapolis, MN. Nov. 2016

International Transportation Economics Association, Santiago, Chile. Jun. 2016

Transportation, Urban, & Regional Seminar, Dept. of Economics, UCI. Nov. 2015

AWARDS & FELLOWSHIPS

North American Regional Science Council

Best Student-Authored Paper Award 2016

Department of Economics, UCI

Summer Research Fellowship 2015–2017

David Brownstone Award for Best Paper in Econometrics 2017

Ken Small Award for Best Paper in Urban & Transportation Economics 2016

AFFILIATIONS

Urban Economics Association; International Transportation Economics Association; The Econometric Society; American Economic Association.

SOFTWARE SKILLS

Matlab; Stata; R; ArcGIS; QGIS; LaTeX; MS Office.

PROFESSIONAL EXPERIENCE

Transport Specialist (JPO), World Bank, Washington, DC. Jul. 2012–Jul. 2013

Corporate Banking Manager, Gulf Bank, Kuwait. Dec. 2007–Jun. 2012

Associate Engineer, Oregon Dept. of Transportation, Salem, OR. Jun.-Sep. 2006

Engineering Intern, Kuwait National Petroleum Co., Kuwait. Jul.-Sep. 2004

OTHER

Vice President, National Union of Kuwaiti Students–USA Branch 2007

Language fluency: Arabic & English.

Citizenship: Kuwait.

ABSTRACT OF THE DISSERTATION

Essays in Urban Economics and Spatial Econometrics

By

Dhari S. R. S. E. Aljutaili

Doctor of Philosophy in Economics

University of California, Irvine, 2018

Professor David Brownstone, Chair

This dissertation features research that contributes to understanding the role of social interactions and social capital in the economy. Social capital (e.g. social norms and networks that facilitate coordination and cooperation in society) has been shown to correlate with desirable socioeconomic outcomes. The three chapters of this dissertation present econometric methods and empirical analyses aimed at evaluating the potential of policy interventions that enhance or leverage the stock of social capital in addressing certain market failures when they emerge.

Chapter 1 studies the relationship between neighborhood design and social capital using a household survey from California. It offers two contributions: (i) an objective measure of social capital–carpooling–that has not been used previously in this context and (ii) a precise definition of the neighborhood using geocoded data. Living on a cul-de-sac (a special case of neighborhood design) is found to be associated with a higher probability of carpooling to school, suggesting that planners can enhance social capital by favoring neighborhood designs that foster social interactions.

Chapter 2 presents a spatiotemporal model for panel count data that preserves the discrete nature of the data, incorporates different forms of dynamics and heterogeneity, accounts for spatial correlation, and is estimated via an efficient Bayesian MCMC algorithm. An empirical

application of solar panel installations in zip codes reveals the presence of spatial correlation that would lead to over-confidence in the estimates if ignored. It also raises a question that is overlooked in the literature about the proper specification of the dynamics.

Chapter 3 is motivated by the well-established result that a market failure in the form of imperfect information causes consumers to under-invest in energy-efficient technologies. It contributes to the literature on peer effects in technology adoption and, hence, whether consumers' social networks can be leveraged by policy makers to fill the information gap and accelerate adoption. Estimating a spatial transition model using individual-level data on solar panel adoption in California and Bayesian MCMC methods, the analysis finds a positive but not statistically important peer effect. The results, though, reveal that failure to control for spatially correlated unobservables leads to biased estimates.

Chapter 1

The Relationship Between Neighborhood Design and Social Capital as Measured by Carpooling

1.1. Introduction

In his popular book–*Bowling Alone: the Collapse and Revival of American Community* (2000), political scientist Robert Putnam blamed what he observed as an eroding social fabric on the decline of *social capital* in American society. In a widely-cited article that inspired the book, Putnam (1995) defined social capital as follows:

> By analogy with notions of physical capital and human capital—tools and training that enhance individual productivity—"social capital" refers to features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit.

Extensive research across different fields of social science is devoted to uncovering links between social capital and socially desirable outcomes. Putnam (1994) shows its positive correlation with economic development as well as effective and stable democratic government.

Social capital also lowers transaction costs and facilitates voluntary cooperation between members of a community, enabling them to overcome dilemmas of collective action (Ostrom, 1990). There is also a growing literature showing how strong neighborhood social networks improve labor market outcomes for their members (Hellerstein et al. 2014; Hellerstein et al. 2011; Bayer et al. 2008)

Though as described above, social capital takes on the definition of a public good that is prone to inefficient private under-provision. Consider a community whose members invest time and effort in cultivating social bonds and establishing norms of trust and reciprocity among them. A situation may arise whereby a few members *free-ride* by consuming the benefits of these bonds and norms without contributing to them (e.g., only a handful of neighbors are needed to perform neighborhood watch duty, but its security benefit extends to all of the neighbors). The question, then, is: can this market failure be corrected by a policy intervention? Urban planning policy has been proposed as an answer. Putnam (2000) suggests that social capital can be enhanced by urban design innovations, such as mixed-use zoning, pedestrian-friendly streets, and public spaces, that promote social interaction among people. Glaeser (2000) also emphasizes the importance of studying the impact of the physical built environment on social capital.

This paper examines these suggestions empirically by analyzing the relationship between neighborhood design and social capital. Using a transportation survey from California, social capital is proxied here by respondents' decisions to carpool to work or school based on the presumption that people carpool only after having interacted socially and established trust between one another. Such interaction can be thought of as an investment in the stock of social capital between its participants, from which they reap the benefit of carpooling as a form of cooperation and reciprocity. Data from the 2012 California Household Travel Survey (CHTS) on respondents' mode choice for their commute to work or school are reclassified into a binary (carpool or not) dependent variable.

The analysis focuses on the cul-de-sac (henceforth, CDS) as a special case of neighborhood design, identified using geographic information system (GIS) software and restricted access to respondents' geocoded residential locations. Focusing on the CDS is motivated by its unique design features that render it relevant to the development of social capital. Because it is a dead-end street with no through traffic, CDS residents experience less and slower auto traffic that may lead them to perceive it as safe for pedestrians and outdoor play by children (Cao et al., 2008). Moreover, persons entering the CDS are more likely to be known to its residents, which increases the probability of spotting strangers, and more importantly criminals, who do not belong in the neighborhood, further enhancing the perceived sense of security and familiarity. These features set the CDS apart from other designs and may serve as potential channels for fostering social interaction. The CDS is convenient for the analysis at hand, but the general policy implication is that urban planners may be able to increase the stock of social capital in a place by favoring spatial designs that foster social interaction between its inhabitants.

The choice of neighborhood design is endogenous though, making it difficult to claim a causal impact on social capital. People who are inherently socially interactive may self-select into a CDS for the very same features mentioned above. Therefore, estimates will be biased unless the neighborhood design effect is disentangled from other confounding effects. Instrumental variables that are unrelated to people's unobserved socialization tendencies but induce an exogenous variation in the choice of neighborhood design are hard to come by since choosing the residential environment is closely related to people's social preferences. Previous attempts at mitigating the self-selection problem involved controlling for people's attitudes and perceptions about the social environment in the neighborhood (Cao et al. 2008 and Kamruzzaman et al. 2014) while others simply ignored the problem altogether. The CHTS data used in the present paper are cross-sectional, thus eliminating the possibility of differencing out unobserved effects as a longitudinal sample would otherwise allow. Also, survey respondents were not asked about their neighborhood preferences. As an al-

ternative approach, the CDS choice and social capital are modeled jointly in this paper in a simultaneous equations model with a rich set of control variables.

The empirical evidence on the relationship between neighborhood design and social capital is mixed and mostly found in the urban planning literature. Cao et al. (2008) report no significant differences between traditional (grid-street) and suburban neighborhoods, but found the preference for CDS to be a significant predictor of children's outdoor play. Kamruzzaman et al. (2014) find that transit-oriented developments and standard suburban neighborhoods, under which the CDS falls according to their definition, are both associated with higher social capital than transit-adjacent developments. Similarly, Wood et al. (2012) and Mason (2010) find higher levels of social capital are associated with CDS-related neighborhoods. Conversely, Mason and Fredericksson (2011) find traditional (grid-street) neighborhoods to be associated with higher social capital. Similarly, Mayo (1979) proclaims that "designers can not directly influence suburban neighboring through street forms" after finding an inverse relationship between CDS and neighbor familiarity and participation. Brown and Cropper (2001), on the other hand, find no significant differences in the sense of community between residents of New Urbanist and standard suburban (which typically include CDS's) neighborhoods.

There is also a small but growing literature in economics on the effect of the built environment in general on social capital. Glaeser and Gottlieb (2006) find mostly negative correlations between density and civic engagement. Borck (2007) find social capital to be positively associated with city size (though not strictly a feature of the built environment). Brueckner and Largey (2008) report a negative effect of density on social interaction, using an instrumental variables approach to address the self-selection problem. Another related, but larger, literature is that on the effect of the built environment on travel behavior. It is relevant to the present paper given the use of a transportation variable (carpooling) as a measure of social capital. Cao and Mokhtarian (2008) and Cao et al. (2009) provide comprehensive surveys of the methodologies and empirical findings of this literature.

This paper offers two main contributions. Carpooling was first introduced as a measure of social capital by Charles and Kline (2006), but, to the best of the author's knowledge, this is its first use in studying the effects of the built environment. To the extent that carpooling is considered both an investment in and a benefit of norms of trust and reciprocity between participants, this argument is even more compelling in the case of carpooling to school; it takes a higher level of trust for people to let their children carpool with those who are otherwise strangers. In that sense, carpooling is an action that embodies the elements of Putnam's definition of social capital. Using this measure, however, is a double-edged sword. On the one hand, it is an advantage in that the survey respondents are blind to the hypothesis at hand. Earlier studies used surveys that elicited respondents' subjective perceptions of their communities' social environments or self-reported level of social interaction, in which there is potential for social desirability bias whereby respondents might provide answers that would portray them as more sociable than they actually were. Using carpooling instead reduces this risk. On the other hand, this measure makes the task of addressing the self-selection problem even harder because it invalidates any instrumental variable candidate that is correlated with respondents' unobserved transportation mode preferences.

The other contribution is the use of a large and comprehensive dataset. While prior studies mostly targeted households in only a handful of neighborhoods, the CHTS sampled 42,431 households across the state of California, allowing the construction of the neighborhood design variable that is representative and includes CDS's of different shapes and lengths. Furthermore, access to the restricted geocoded residential locations allows identifying households who reside on CDS's and how far from the end of the CDS they are. It also helps in overcoming a major challenge often faced by urban economic researchers who attempt to analyze data from large surveys. Because of privacy concerns, exact locational data are often unavailable, forcing researchers to resort to coarser definitions of the neighborhood (e.g. census tract or zip code). This is not an issue in this paper.

1.2. Data

The CHTS data are obtained from the National Renewable Energy Laboratory (NREL).[1] It is a statewide survey lead by the California Department of Transportation for transportation and environmental policy and planning purposes. It sampled 42,431 households (encompassing 109,113 individuals) from all of California's 58 counties. In addition to socioeconomic information, households recorded their travel activities in a diary for a pre-assigned 24-hour period, covering every day of a full year, and some were given global position system (GPS) devices. The data were retrieved from the travel diaries, GPS devices, computer-assisted telephone interviews, and online forms.

GIS software is used to identify CDS streets, which vary in shape and length, and households' positions along them also vary. Thus, some criteria must be established to form the basis upon which a household is considered as belonging to a CDS neighborhood. Ben-Joseph (1995) reports that the Institute of Transportation Engineers recommends 1000ft as the maximum length of the CDS street, but in a survey of 75 U.S. city planning codes (including 56 from California), he found a majority of them set the maximum length between 500 and 600ft (approximately 152 and 183m). In this analysis, 150m is chosen as a reasonable length. However, households may experience different levels of the neighborhood social environment depending on where along the CDS street they are located (e.g., parents may be more willing to let their children play outdoor if their home is directly at the circular end of the CDS than those located at its other end). This possibility will be examined in a sensitivity analysis by varying the CDS length (50, 100, and 200m). The second criterion is that a CDS has only one point of access such that a resident can only enter and exit the CDS through that point (equivalently, an individual must be able to reach the circular end of the CDS from their home without having to cross a street). Finally, while the automated

[1] "Transportation Secure Data Center" (2015). National Renewable Energy Laboratory. October 25^{th}, 2015. www.nrel.gov/tsdc.

GIS algorithm does not differentiate between a typical CDS and a standard dead-end street, instances of the latter are manually re-coded as non-CDS[2].

The binary dependent variable (carpooling) is constructed from the respondents' reported mode of transportation to work and school out of an exhaustive list of 29 modes. It is worth noting that "auto passenger" and "carpool/vanpool" are listed as separate modes. Responding to an e-mail enquiry, the survey manager[3] explained that unless respondents asked for additional clarification, it was assumed that they would consider the latter to be travel with non-relatives while the former signifies travel with relatives. "Private shuttle (employer)" is also listed as a mode, which rules out group transportation arranged by employers from being considered as resulting from social interaction.

One crucial limitation in the data is that they do not indicate that carpooling is necessarily occurring between individuals in the same neighborhood, which could undermine the ability to attribute it to the neighborhood design or the social environment as other factors may be responsible for it. For example, a worker may pick up a co-worker who lives on their driving path, indicating that carpooling is a result of proximity or social interaction at the job location. The presence of amenities shared by several neighborhoods may also provide opportunities for social interaction. For example, children from different but proximate neighborhoods may form social bonds by playing at a public park nearby. In this case, carpooling between them would be attributed to the park. In this paper, the contrast between the results from the work and school analyses helps gain some insight about the patterns of carpooling activity, but without specific information on who is carpooling with whom, the results should be interpreted with caution.

[2]A previous version of the paper did not differentiate between CDS's and standard dead-end streets based on the assumption that, as they relate to social capital, they are functionally similar. However, upon close examination, this was not found to be the case; many streets identified by the automated GIS algorithm as dead-ends are in fact intersection nodes in urban areas or streets that lead to remote properties in rural areas. Neither case constitutes a neighborhood comparable to a typical CDS. Hence, the analysis is restricted to CDS's while standard dead-ends are removed.

[3]Daigler, Vivian (NuStats Research Solutions). "Re: CHTS 2012." Received on 24 Aug. 2017.

To control for destination accessibility, travel times of the shortest paths (in terms of driving time) from home to work and school are calculated based on speed limits, using work and school geocoded locations. In some instances, the travel times are found to be unusually large because the sample included individuals working or studying in different regions or states (despite the survey manual's specific instructions to exclude them). To address this problem, only travel times less than 90 minutes are included in the analysis.

The survey oversampled "hard-to-reach" groups based on demographic (hispanic, low-income, young, or large households) and transit-use (households with zero-vehicles or residing near transit facilities) factors. Table 1.1 shows weighted and unweighted summary statistics of the sample of households, which do not show any appreciable differences. Therefore, the unweighted estimates will be reported as the paper's main results.

Variables with "don't know" or "refuse to answer" values, which represent a negligible share of the sample, are treated as missing and excluded from the analysis. One exception is made for the household income variable, for which these missing values constitute 8.6% of the sample. A "missing income" category is added to the income categories.

Finally, supplementary variables were constructed using data from the following secondary sources: 2010 U.S. Census (population density and dissimilarity indices), 2012 American Community Survey 5-year estimates (zip code variables), and the California Department of Justice (2012 murder rate).

Summary Statistics

Table 1.1 provides summary statistics of a subset of household characteristics broken down by CDS status. Tests of mean difference (for continuous variables) and independence (for categorical variables) show a statistically significant difference in most characteristics based on CDS status. Most importantly, households tend to carpool to school or work more if they reside in CDS neighborhoods. Travel time to work is statistically significantly longer for CDS households, which is consistent with the fact that CDS neighborhoods tend to be in

Table 1.1: Household summary statistics ($N = 42,431$)

	CDS ($N = 3,264$) Mean (SD)	Non-CDS ($N = 39,167$) Mean (SD)	Mean Comparison Test Stat.[b]	Full Sample Unweighted Mean (SD)	Full Sample Weighted Mean (SD)
Endogenous variables					
Census tract density[a]	4.793 (3.729)	6.463 (8.283)	11.4	6.335 (8.037)	7.693 (8.868)
Carpool to work[c]	0.028	0.017	20.4	0.018	0.017
Carpool to school[d]	0.030	0.011	88.4	0.013	0.015
Household characteristics					
Household size	3.088 (1.356)	2.529 (1.367)	-22.5	2.572 (1.374)	2.695 (1.501)
No. of children $\leq$ 16 years	0.716 (1.056)	0.461 (0.915)	-15.1	0.481 (0.929)	0.556 (0.989)
No. of students $\leq$ 8th grade	0.422 (0.785)	0.268 (0.653)	-12.7	0.280 (0.665)	0.326 (0.711)
No. of workers	1.695 (0.735)	1.182 (0.883)	-32.2	1.222 (0.883)	1.186 (0.944)
No. of vehicles	2.286 (0.936)	1.827 (0.993)	-25.5	1.862 (0.997)	1.812 (1.065)
Tenure (years)	15.21 (10.37)	15.99 (12.74)	3.42	15.93 (12.58)	15.75 (12.78)
Non-white householder	0.217	0.262	33.4	0.258	0.292
Hispanic householder	0.142	0.186	39.6	0.182	0.211
Unemployed householder	0.029	0.040	11.1	0.039	0.048
Householder college degree	0.596	0.495	123	0.503	0.477
Income(2) $25,000 - $49,999	0.113	0.193	127	0.187	0.221
Income(3) $50,000 - $99,999	0.327	0.299	11.0	0.301	0.264
Income(4) $100,000 - $149,999	0.231	0.146	169	0.153	0.110
Income(5) $150,000 or more	0.216	0.115	291	0.122	0.126
Income(6) Income missing	0.070	0.087	12.0	0.086	0.078
Avg. travel time to work [c]	20.37 (14.04)	19.16 (15.07)	-4.30	19.30 (14.96)	18.48 (13.83)
Avg. travel time to School[d]	7.415 (7.162)	7.257 (7.971)	-0.544	7.276 (7.875)	6.867 (7.154)

Variables without a standard deviation are binary.
[a] Census tract population density in units of 1000 people per square-mile.
[b] t-test of mean difference for continuous variables or chi-squared test of independence for categorical variables.
[c] For the subset of households with one or more non-home-based workers ($n = 26,781$).
[d] For the subset of households with one or more non-home-based students from kindergarten to 8th grade ($n = 6,777$).

the suburbs away from the city center where employment is normally concentrated. School distance, on the other hand, shows no statistically significant difference, reflecting the fact that schools are more evenly distributed over space. Overall, households tend to reside closer to their schools than their employment locations.

Table 1.2 shows summary statistics for the subsamples of individual workers and students used in the analysis. The reported values indicate which variables were included in each of the work and school models. The dissimilarity indices are used to control for the possibility that individuals may be more socially interactive with others of the same race or ethnicity. Dissimilarity is a measure of segregation between two groups in a geographical area. Racial

Table 1.2: Individual summary statistics

Subsample	Workers ($N = 38,494$) Mean (SD)[a]	Students ($N = 10,324$) Mean (SD)[a]
Endogenous variables		
Reside in CDS neighborhood	0.125	0.124
Census tract density	6.357 (7.527)	6.267 (6.869)
Carpool	0.019	0.027
Individual characteristics		
Age	47.36 (13.49)	9.371 (2.599)
Male	0.519	0.512
Black	0.028	0.028
American Indian or Alaskan Native	0.046	0.074
Asian	0.068	0.063
Mixed or other race	0.128	0.230
Hispanic	0.212	0.412
Citizen or in US 10+ years	0.982	0.970
Education: college or higher	0.509	-
Travel time (minute)	19.15 (16.06)	6.104 (7.577)
Household characteristics		
Age of householder	50.99 (12.17)	42.92 (9.404)
Male householder	0.473	0.418
Household size	3.061 (1.414)	4.629 (1.213)
Number of vehicles	2.202 (1.015)	1.981 (0.841)
Tenure (years)	14.42 (10.88)	9.213 (7.549)
Number of children	0.632 (0.997)	2.353 (0.985)
1-worker household	0.342	0.447
2-or-more-workers household	0.658	0.506
Number of students	0.359 (0.727)	1.831 (0.804)
Householder lives w/ spouse/partner	0.764	0.852
Householder unemployed	0.023	0.048
Householder retired or homemaker	0.121	0.199
Householder other employment status	0.031	0.050
Retired or homemaker in household	0.165	0.337
Householder education: college or higher	0.543	0.462
Income(2) $25,000 - $49,999	0.147	0.175
Income(3) $50,000 - $99,999	0.329	0.271
Income(4) $100,000 - $149,999	0.205	0.173
Income(5) $150,000 or more	0.181	0.163
Income missing(6)	0.062	0.045
Avg. household travel time to school (minute)	2.937 (6.365)	-
Avg. household travel time to work (minute)	-	16.93 (15.55)
County characteristics		
Racial dissimilarity (white vs. non-white)	0.309 (0.066)	0.304 (0.066)
Ethnic dissimilarity (hispanic vs. non-hispanic)	0.402 (0.104)	0.402 (0.100)
Murder rate	4.971 (3.220)	5.029 (3.143)
Regions		
(2) Central Coast	0.052	0.042
(3) Central Sierra	0.019	0.013
(4) Greater Sacramento	0.054	0.057
(5) Northern California	0.035	0.026
(6) Northern Sacramento Valley	0.021	0.022
(7) San Joaquin Valley	0.123	0.159
(8) Southern Border	0.054	0.061
(9) Southern California	0.375	0.382
Instrument		
% pre–1940 homes in county subdivision	0.091 (0.105)	0.083 (0.096)

[a] Variables without a standard deviation are binary, equaling 1 if the condition is met and 0 otherwise.

(white vs. non-white) and ethnic (hispanic vs. non-hispanic) dissimilarity indices at the county level are calculated using a formula proposed by Duncan and Duncan (1955). Residing in a high dissimilarity area is correlated with being more exposed to people of own race or ethnicity.

1.3. Methodology

Since this is the first use of carpooling as a measure of social capital in the current context, it is prudent to ask how well does it perform in comparison to the more direct measures used in prior studies? Ideally, one way to answer this question is by replicating one of those studies but with using the carpooling variable and comparing the results. However, an exact replication is not possible since none of the prior studies had data on carpooling. As a second best, a quasi-replication exercise of a prior study (i.e., replicating the model to answer the same question but with using the CHTS data and the carpooling dependent variable) can be performed and results can be compared qualitatively. The Brueckner and Largey (2008) study cited earlier (henceforth, BL2008) is used for this exercise as a first pass to judge the reliability of using carpooling as a measure of social capital.

Following this exercise, the main model of the relationship between neighborhood design and social capital is presented and estimated.

1.3.1. Replication Exercise

Summary of BL2008

The BL2008 study used national cross-sectional data from the 2000 Social Capital Benchmark Survey to estimate the effect of census tract population density on 10 social interaction measures involving interacting and cooperating with neighbors, confiding with people, socializing with friends, and membership in hobby-based clubs and non-church groups. The authors employed an instrumental variables approach to address the problem of self-selection

into census tracts. Population densities of metropolitan statistical areas (MSA) and urban areas were used as instruments for census tract density. Their identifying assumption was that respondents' choose their metro area based on factors (such as job location and family ties) that are unrelated to their tendencies to be socially interactive. They also use MSA terrain ruggedness[4] as an additional instrument based on the assumption that residential development is constrained by terrain ruggedness, thus affecting density.

For the binary dependent variables, a joint model of density and social interaction was estimated by maximum likelihood. For the continuous dependent variables, a two-stage least-squares model was estimated. Nine of the 10 estimations produced negative coefficients for density, 7 of which were statistically significant (the 10^{th} was positive but not statistically significant). The authors concluded that the evidence did not support the claim that urban sprawl erodes social capital. To the contrary, their results indicated that, controlling for self-selection, people living in the low-density suburbs are more socially interactive than those in city centers. They also showed estimates from a single-equation model that ignores the endogeneity were biased upward, taking it as evidence of self-selection by respondents into census tracts with higher density.

Replication of BL2008

Mimicking BL2008's analysis, the effect of census tract density on social capital is estimated but using the CHTS data and carpooling as the dependent variable instead. Two separate models for carpooling to work and school are estimated by maximum likelihood, using the same instrumental variables as in BL2008 and a similar set of covariates augmented by additional ones relevant to the current setting. In each model, a 2-equation system is used to jointly model the binary outcome (carpooling) and the continuous endogenous variable (density).

[4]Obtained from Burchfield et al. (2006).

To evaluate the performance of the instruments, first-stage results are obtained by re-estimating the models via two-stage least-squares, which reveals that the MSA terrain ruggedness is a weak instrument in the school model but relevant in the work model. Therefore, the ruggedness instrument is removed from the school model in the replication exercise. Summary statistics of the estimation samples reveal that the students subsample has a lower and less variable average census tract density (7,034 persons per sqm. with a standard deviation of 6,976) than the workers subsample (7,276 persons per sqm. with a standard deviation of 7,712). Coupled with the fact that there are far fewer students than workers in the data, the weakness of the terrain ruggedness instrument in the school model might be attributed to the fact there is less variability in the density for it to induce. The discrepancy in the instrument's performance as it relates to BL2008 may also be due to their use of a national sample with sufficient variation in terrain as opposed to only California.

Replication Results

Table 1.3 presents the average partial effects (APE) of census tract population density on carpooling to work and school along with results of statistical tests. The complete first-stage and reduced-form results are presented in appendix A.

Table 1.3: Replication results: effect of census tract density on carpooling

	Dependent variable	
	Carpool to work	Carpool to school
APE of census tract density	-0.0003 (0.0004)	-0.0030 (0.0012)**
Correlation, ρ	0.028	0.188
Wald exogeneity ($H_o : \rho = 0$) test statistic [P-value]	0.20 [0.652]	4.23 [0.040]
N	33,140	9,581
Number of clusters	22,838	6,264
First-stage[a] F-statistic	1041	436.0
APE from single-equation probit model	-0.0001 (0.0001)	-0.0008 (0.0004)*

Statistical significance: $*\,(P < 0.10)$; $**\,(P < 0.05)$; $***\,(P < 0.01)$
Standard errors are in parentheses; P-values are in square brackets.
[a] First-stage is obtained by re-estimating the model with a linear two-stage least-squares estimator.

In both models, density has a negative effect on carpooling to work and school, but only the latter is statistically significant, implying that students in lower-density census tracts

tend to be more socially interactive as measured by their carpooling to school. This result is consistent with the findings of BL2008. Moreover, the bottom row of Table 1.3 reports estimate from a single-equation probit model, that is, ignoring the endogeneity of density. These estimates are biased upward compared to those from the 2-equation joint model, implying the error term is positively correlated with density. In other words, it suggests that individuals who are inherently socially interactive tend to reside in denser census tracts. This observation about self-selection behavior is also consistent with the one made by BL2008.

The lack of statistical significance for the effect on carpooling to work also sheds some light on the performance of the carpooling variable as a measure of social capital. As mentioned previously, the CHTS data do not indicate that carpooling is occurring exclusively between neighbors. In the workers case, carpooling might be a result of factors related to the destination, such as proximity or social interaction at the job location, and not necessarily due to density at home. On the other hand, it is more likely that students who live in the same census tract also attend the same or nearby school (Table 1.1 showed that, on average, households reside closer to their schools than their job locations). In other words, finding a carpool partner who shares the destination location is less of an obstacle for a student than it is for a worker. Therefore, attributing the estimated effect to density at home is more plausible in the school model than the work model.

Based on this intuition, the contrast in statistical significance between the two models shows that the observed carpooling variable operates in the way social interaction is expected to operate. This conclusion and the qualitative agreement with the findings of BL2008—and other studies in the literature that showed the adverse effects of low density on social capital disappearing after accounting for self-selection—provide the key take-away from the replication: the results do not rule out carpooling as a plausible measure of social capital.

1.3.2. Model of Neighborhood Design and Social Capital

For the main analysis, let D_i denote neighborhood design, equaling 1 if individual i resides on a CDS and zero otherwise. Let S_i denote social capital, equaling 1 if the individual carpools and zero otherwise. These two decisions are modeled jointly as follows:

$$D_i = 1\{\boldsymbol{x'_{1i}}\boldsymbol{\alpha} + \varepsilon_{1i} > 0\} \tag{1.1}$$

$$S_i = 1\{\boldsymbol{x'_{2i}}\boldsymbol{\beta} + D_i\gamma + \varepsilon_{2i} > 0\}, \tag{1.2}$$

where $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are vectors of (common and unique) exogenous variables, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of their associated coefficients, and γ is the coefficient of interest associated with neighborhood design. The errors account for unobserved determinants of D_i and S_i and are assumed to be distributed bivariate normal:

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \Bigg| \, \boldsymbol{x_{1i}}, \boldsymbol{x_{2i}} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

where ρ is the correlation between the unobservables and the unit-variances are the usual normalization in probit models. This model is commonly known as the *recursive bivariate probit* in that D influences S, but not vice-versa (Maddala, 1983, p.122). The coefficients, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \rho)$, are estimated by maximizing the log-likelihood function.

The neighborhood design equation (1.1) is identified by construction since it is determined by exogenous variables. Identification of the social capital equation (1.2) is possible theoretically by relying solely on the model's non-linearity (Wilde, 2000); however, Wooldridge (2010, p.596-599) showed how this approach sometimes results in poor convergence. Instead, identification can be improved using variables that appear in $\boldsymbol{x_1}$ but not in $\boldsymbol{x_2}$, and thus are excluded from (1.2). In instrumental variable terminology, these are variables that are relevant (correlated with D_i), exogenous (uncorrelated with ε_{2i}), and excludable (influence S_i only through D_i).

For this analysis, the fraction of housing units built prior to the year 1940 at the county subdivision level is used as an instrument. The relevance of this variable is based on the historical argument that the urban sprawl phenomenon, of which the CDS is a prominent feature, coincided with the post-World War 2 economic expansion. As such, a subdivision with a higher fraction of units built prior to the war is less likely to contain CDS neighborhoods. Thus, the variable is expected to be negatively correlated with D_i. To satisfy the exogeneity condition, recall that the variable of interest here—neighborhood design—represents a choice at a much finer spatial scale (i.e., street level) than the subdivision. In California, county subdivisions are areas delineated by the Census Bureau for statistical purposes. There are 58 counties in California divided into 397 subdivisions (e.g., Los Angeles county has 20 subdivisions while San Diego county has 13). The choice of this spatial scale for the instrument is reasonable given the fine scale of D_i. Otherwise, an instrument at, say, the census tract or zip code levels would likely violate the exogeneity condition, whereas an instrument at a much coarser scale, such as county or MSA levels, would lose its relevance to D_i. The key identifying assumption is that a household first chooses the general area of the county (i.e., county subdivision) to move to based on factors unrelated to social capital or preference for carpooling (e.g., density, amenities, terrain, etc.), and then chooses the neighborhood design within their chosen area. Thus, to the extent that this sequential decision process is realistic, the choice of county subdivision is assumed to be pre-determined outside the model. The assumption would be violated if a household first developed a preference for a CDS neighborhood and then chose a subdivision made up primarily of CDS's. This reversal of the decision sequence is plausible, but throughout the manual GIS procedure conducted earlier to identify CDS households, areas made up primarily of CDS's were extremely rare.

Partial Effects

As with any non-linear model, coefficient estimates from the bivariate probit model are not interpretable beyond their signs. Partial effects must be computed in order to produce

meaningful results. Of primary interest is the partial effect of changing the endogenous neighborhood design variable, D, on the probability of carpooling.

There are different kinds of partial effects that can be computed. One of which is the effect on the *marginal* probability of carpooling:

$$E(S \mid \boldsymbol{x_2}, \rho = 0, D = 1) - E(S \mid \boldsymbol{x_2}, \rho = 0, D = 0) = \Phi(\boldsymbol{x_2'\beta} + \gamma) - \Phi(\boldsymbol{x_2'\beta}), \tag{1.3}$$

where $\Phi(.)$ denotes the cumulative distribution function (cdf) of the univariate standard normal distribution. Another type of partial effect is that on the *conditional* probability of carpooling:

$$E(S \mid \boldsymbol{x}, \rho, D = 1) - E(S \mid \boldsymbol{x}, \rho, D = 0) = \frac{\Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho)}{\Phi(\boldsymbol{x_1'\alpha})} - \frac{\Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho)}{1 - \Phi(\boldsymbol{x_1'\alpha})} \tag{1.4}$$

where $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$ and $\Phi_2(.)$ denotes the bivariate standard normal cdf.

Which partial effect is appropriate? It depends on the policy experiment of interest. First notice that if the correlation (ρ) is zero, (1.4) collapses to (1.3), which highlights the difference in interpretation between the two. The effect on the marginal probability (1.3) represents a ceteris-paribus scenario whereby the policy experiment involves changing the household's CDS status while holding all other *observed and unobserved* variables constant. On the other hand, the effect on the conditional probability (1.4) holds *only the observed* variables constant, keeping ρ unrestricted to account for changes in the unobserved variables as a result of the policy experiment. As such, (1.4) is appropriate if the policy maker is interested not only in the effect on social capital, but also on how other effects on unobserved variables might interfere with it, either by amplifying or attenuating it, depending on the sign of ρ. In the results section, both effects (evaluated at each observation and averaged over the sample) will be reported and discussed. For convenience, the estimate using equation (1.3)

will be referred to as the *Marginal APE* while that from equation (1.4) will be referred to as the *Conditional APE*. Standard errors for the partial effects are calculated using the delta method.

The complete derivation of the joint probabilities, log-likelihood function, and the partial effects of all variables is presented in appendix A.

1.4. Estimation Results

Two separate models are estimated: the first examines the effect of neighborhood design on carpooling to work using the subsample of individual non-home-based workers and the second examines the effect on carpooling to school using the subsample of non-home-based school students between Kindergarten and 8th grade (High school students are excluded because, since the driving age in California is 16, they might carpool with schoolmates with whom they formed relationships at school and not as a result of the neighborhood environment).

1.4.1. Effect of Neighborhood Design on Social Capital

Tables 1.4 and 1.5 present the estimated APE of living in a CDS neighborhood on social capital as measured by carpooling to work and school, respectively. The CDS coefficient (γ) as well as the marginal and conditional APE's are reported for several model specifications in columns (1) to (5) based on which covariate groups are included as indicated in the table, with (5) being the preferred specification. In addition, column (6) presents estimates from a restricted model specification similar to (5) but with the correlation fixed at zero (i.e., equivalent to a univariate probit that does not account for the endogeneity). The lower two panels of the tables report the correlation estimates, diagnostic statistical tests, and first-stage results from re-estimating the models using a two-stage least-squares estimator.

Note that under the restricted model, the two types of APE are numerically equivalent, but the estimate is reported on the row pertaining to the conditional APE for conceptual consistency. To elaborate, recall that the marginal APE imposes the restriction $\rho = 0$ regardless of the estimated ρ recovered along with the coefficients from the estimation. On the other hand, the conditional APE maintains ρ from the model estimation as is in computing the partial effects. Since under the restricted model both the coefficient estimates and the partial effects are based on $\rho = 0$ by construction, the resulting APE is thus conceptually comparable to the conditional APE's under specifications (1)-(5).

Work Model

Table 1.4 shows the results for the model with carpooling to work as the dependent variable. With the exception of the first specification that does not include any covariates besides the instrument, none of the coefficient and APE estimates is statistically distinguishable from zero. The first-stage results indicate the instrument employed here is relevant, with F-statistics well above the rule of thumb of 10. The Wald exogeneity tests for specifications (2) to (5) fail to reject the null hypothesis that $\rho = 0$, implying the D is not endogenous. As such, more efficient estimates can be obtained by estimating the model with ρ restricted to zero. Indeed, the APE estimate in column (6) has a smaller standard error, but it does not change the conclusion that living in a CDS neighborhood has no appreciable effect on social capital as measured by carpooling to work. This result is consistent with the discussion of the replication results; the decision to carpool to work is likely more attributable to factors related to the workers' destinations (e.g. proximity or social interaction at the job location) than the social environment in their residential neighborhood.

School Model

The results from the school model are presented in Table 1.5. All of the APE estimates are positive, but only the marginal APE under specification (2) is statistically significant at the 10% level. The magnitudes of the effects get smaller as more covariate groups are

Table 1.4: Estimates of the effect of CDS on carpooling to work

	Unrestricted ρ					Restricted ρ
Specification	(1)	(2)	(3)	(4)	(5)	(6)
Coefficient, γ	0.9259	−0.0859	0.1391	0.0831	0.0072	−0.0233
	(0.4169)**	(0.3624)	(0.3608)	(0.2838)	(0.3038)	0.0521
Marginal APE (eq. 1.3)	0.0992	−0.0036	0.0069	0.0039	0.0003	
	(0.0807)	(0.0143)	(0.0198)	(0.0142)	(0.0136)	
Conditional APE (eq. 1.4)	−0.0020	−0.0009	−0.0010	−0.0009	−0.0011	−0.0010
	(0.0195)	(0.0156)	(0.0154)	(0.0122)	(0.0129)	(0.0022)
Covariate group						
Instrument	✓	✓	✓	✓	✓	✓
Individual		✓	✓	✓	✓	✓
Household			✓	✓	✓	✓
County				✓	✓	✓
Regional fixed effects					✓	✓
Correlation, ρ	−0.44	−0.03	−0.08	−0.05	−0.02	0.00
Wald exogeneity test of $\rho = 0$	5.12	0.03	0.18	0.12	0.01	
	[0.02]	[0.87]	[0.67]	[0.73]	[0.92]	
Goodness of fit: Wald $\chi 2$ (d.f.)	265 (2)	882 (22)	1164 (55)	1225 (61)	1338 (77)	1339 (77)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
N	38,607	38,607	38,485	38,485	38,485	38,485
Clusters	26,772	26,772	26,668	26,668	26,668	26,668
First-stage[a] F-statistic	877	880	79.0	76.5	60.0	
First-stage[a] adjusted R^2	0.02	0.02	0.03	0.04	0.04	

Statistical significance: $*\,(P < 0.10)$; $**\,(P < 0.05)$; $***\,(P < 0.01)$
Standard errors are in parentheses; P-values are in square brackets.
[a] First-stage is obtained by re-estimating the model with a linear two-stage least-squares estimator.

added and soak up the variation in carpooling. The conditional APE's are similar across all specifications since the effect of adding covariates on the APE estimate is offset by the unrestricted correlation. The Wald test fails to reject the exogeneity of D, indicating the model can be estimated more efficiently by restricting ρ at zero. Estimates from the restricted model (column 6) show a positive and statistically significant APE, indicating that, on average, residing in a CDS neighborhood is associated with an increase of 1.5 percentage points in the probability of carpooling to school.

To evaluate the economic significance of the effect, it must be compared to the sample average predicted probability of carpooling, which is given by the conditional mean function

$$\begin{aligned} E(S \mid \boldsymbol{x}) &= \hat{\Pr}(S = 1, D = 1) + \hat{\Pr}(S = 1, D = 0) \\ &= \Phi_2(\boldsymbol{x_2'}\hat{\boldsymbol{\beta}} + \hat{\gamma}, \boldsymbol{x_1'}\hat{\boldsymbol{\alpha}}, \hat{\rho}) + \Phi_2(\boldsymbol{x_2'}\hat{\boldsymbol{\beta}}, -\boldsymbol{x_1'}\hat{\boldsymbol{\alpha}}, -\hat{\rho}), \end{aligned} \tag{1.5}$$

Table 1.5: Estimates of the effect of CDS on carpooling to school

	Unrestricted ρ					Restricted ρ
Specification	(1)	(2)	(3)	(4)	(5)	(6)
Coefficient, γ	1.954	1.675	0.8504	0.7342	0.4335	0.2249
	(0.8106)**	(0.4876)***	(0.5247)	(0.5590)	(0.6021)	(0.0851)***
MarginalAPE (eq. 1.3)	0.4473	0.3219	0.0957	0.0715	0.0335	
	(0.3220)	(0.1753)*	(0.0950)	(0.0850)	(0.0613)	
Conditional APE (eq. 1.4).	0.0197	0.0144	0.0128	0.0127	0.0138	0.0150
	(0.1014)	(0.0476)	(0.0410)	(0.0428)	(0.0456)	(0.0065)**
Covariate group						
Instrument	✓	✓	✓	✓	✓	✓
Individual		✓	✓	✓	✓	✓
Household			✓	✓	✓	✓
County				✓	✓	✓
Regional fixed effects					✓	✓
Correlation, ρ	−0.72	−0.66	−0.33	−0.27	−0.11	0.00
Wald exogeneity test of $\rho = 0$	2.59	6.25	1.454	0.86	0.13	
	[0.11]	[0.01]	[0.21]	[0.35]	[0.72]	
Goodness of fit: Wald $\chi 2$ (d.f.)	79.3 (2)	269 (18)	420 (56)	421 (62)	421 (77)	425 (78)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
N	10,324	10,324	10,324	10,324	10,324	10,324
Clusters	6,775	6,775	6,775	6,775	6,775	6,775
First-stage[a] F-stat.	194	188	31.5	28.0	23.7	
First-stage[a] adjusted R^2	0.01	0.02	0.04	0.04	0.05	

Statistical significance: $*$ $(P < 0.10)$; $**$ $(P < 0.05)$; $***$ $(P < 0.01)$
Standard errors are in parentheses; P-values are in square brackets.
[a] First-stage is obtained by re-estimating the model with a linear two-stage least-squares estimator.

which is calculated to be 2.7%, reflecting the fact from Table 1.2 that only a very small percentage of the sample carpools to school. Therefore, while the estimated 1.5-percentage-point increase in the probability seems small in the abstract, it is nonetheless substantial relative to the sample average probability.

1.4.2. Exogenous Covariates

Only the estimates from the school model are reported and discussed here (since the work model does not produce statistically significant APE's on carpooling, the effects of exogenous covariates are not of interest and not discussed here. They are, nevertheless, tabulated in appendix A for completeness). Appendix A contains a description of the covariates and the rationale behind including them in the model.

Table 1.6 presents the reduced-form estimates of the APE's of the covariates included in specification (6). Since neighborhood design is assumed determined by exogenous variables only, the reported total effects (first two columns) represent the direct effect on D and are equivalent to what would have been obtained from a univariate probit model with D as the dependent variable. On the other hand, the reported total effects on carpooling to school (last two columns) equal the sum of the direct (on S) and indirect (on S through D) effects.

Neighborhood Design Equation (1.1)

The effects on the choice of neighborhood design (CDS or not) and their standard errors are presented in the first two columns. The effect of the instrumental variable (fraction of housing units in a county subdivision built prior to 1940) is statistically significant and has the expected negative sign. Ethnicity does not have a statistically significant association with neighborhood design whereas out of the race categories, only being black is negatively associated with CDS relative to being white. Households headed by older males tend to reside in CDS neighborhoods. Larger households are negatively associated with CDS whereas those with more children or two or more workers tend to reside in CDS neighborhoods. Household income and the number of vehicles are also positively associated with CDS. Travel time to school is not statistically significant, but CDS households tend to have a longer average travel time to work. None of the county-level variables (dissimilarity and murder rate) is statistically significant while some of the regional dummy variables are.

Social Capital Equation (1.2)

The total effects of the exogenous variables on social capital as measured by carpooling to school are reported in the last two column. Of the individual characteristics, students who carpool tend to be older and reside farther (in terms of travel time) from school. Black, Asian, and mixed-race students tend to carpool less relative to white students. Hispanic students also carpool less. Of the household characteristic, only the householder having a college degree or higher has a (positive) statistically significant effect at the 5% level. Other effects

that are statistically significant only at the 10% level are the number of children (positive), the $100,000-$149,999 income category (positive), having a male householder (negative), and the presence of a retired person or homemaker aged 18-65 in the household (negative). County racial dissimilarity has a positive and statistically significant effect, indicating that students who are more exposed to people from their own race in their residential environment are more likely to carpool to school. Finally, the indirect effect of the instrument (fraction of homes built prior to 1940) is negative and statistically significant.

Table 1.6: Effects of exogenous variables in the restricted ($\rho = 0$) school model

Explanatory variables	**Dependent variables**			
	Neighborhood design (D)		Carpool to school (S)	
	APE	(SE)	APE	(SE)
Individual Characteristics				
Age	-	-	0.0028	(0.0008)***
Male	-	-	−0.0025	(0.0031)
Black[a]	−0.0403	(0.0224)*	−0.0165	(0.0063)**
American Indian or Alaskan Native [a]	−0.0158	(0.0198)	0.0012	(0.0084)
Asian[a]	−0.0088	(0.0163)	−0.0137	(0.0054)**
Mixed or other race[a]	−0.0003	(0.0118)	0.0107	(0.0065)*
Hispanic	−0.0060	(0.0114)	−0.0145	(0.0051)***
Citizen or in US 10+ years	-	-	0.0086	(0.0098)
Travel time to school	−0.0003	(0.0005)	0.0008	(0.0002)***
Household Characteristics				
Household size	−0.0120	(0.0070)*	−0.0002	(0.0001)
Number of children	0.0206	(0.0088)**	0.0003	(0.0002)*
Number of students	-	-	−0.0030	(0.0032)
1-worker household	0.0331	(0.0219)	0.0013	(0.0092)
2-or-more-workers household	0.0488	(0.0239)**	0.0058	(0.0099)
Income(2)[a] $25,000 - $49,999	0.0234	(0.0134)*	−0.0034	(0.0068)
Income(3)[a] $50,000 - $99,999	0.0470	(0.0138)***	0.0016	(0.0075)
Income(4)[a] $100,000 - $149,999	0.0832	(0.0171)***	0.0156	(0.0093)*
Income(5)[a] $150,000 or more	0.1035	(0.0199)***	0.0096	(0.0095)
Income missing[a]	0.0650	(0.0226)***	0.0056	(0.0100)
Tenure (years)	0.0007	(0.0006)	0.0004	(0.0003)
Number of vehicles	0.0105	(0.0063)*	−0.0040	(0.0030)
Household average travel time to work	0.0005	(0.0003)*	0.0000	(0.0001)
Age of householder	0.0010	(0.0006)*	0.0001	(0.0003)
Male householder	0.0179	(0.0097)*	−0.0066	(0.0039)*
Householder lives w/ spouse/partner	−0.0005	(0.0147)	0.0038	(0.0058)
Householder has college degree or higher	0.0155	(0.0109)	0.0090	(0.0045)**
Householder unemployed[a]	0.0100	(0.0253)	0.0002	(0.0004)
Householder retired or homemaker[a]	0.0178	(0.0153)	0.0003	(0.0003)
Householder other employment status[a]	0.0347	(0.0277)	0.0006	(0.0005)
Retired or homemaker aged 18-65 in household	-	-	−0.0086	(0.0050)*
County characteristics				
Racial dissimilarity	0.0789	(0.0970)	0.1040	(0.0487)**
Ethnic dissimilarity	−0.0983	(0.0810)	−0.0368	(0.0360)
Murder rate	0.0011	(0.0016)	−0.0008	(0.0006)
Regional dummy variables[a]				
(2) Central Coast	−0.0420	(0.0220)*	0.0072	(0.0115)
(3) Central Sierra	−0.0691	(0.0330)**	−0.0019	(0.0135)
(4) Greater Sacramento	−0.0134	(0.0218)	0.0079	(0.0098)
(5) Northern California	−0.0921	(0.0240)***	−0.0097	(0.0086)
(6) Northern Sacramento Valley	−0.0032	(0.0365)	−0.0049	(0.0102)
(7) San Joaquin Valley	−0.0171	(0.0180)	−0.0008	(0.0065)
(8) Southern Border	0.0177	(0.0234)	0.0054	(0.0092)
(9) Southern California	−0.0152	(0.0148)	0.0112	(0.0066)*
Instrument				
% pre-1940 housing units in county subdivision	−0.5822	(0.0699)***	−0.0098	(0.0044)**

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)

[a] Reference categories: Income ($< \$25,000$); race (white); employment status (employed); region (Bay Area)

1.4.3. Sensitivity Analysis

CDS Street Length

Table 1.7 examines the sensitivity of the estimate to varying the length of the CDS street, using the preferred model specification (5) that includes all covariate groups as well as regional dummy variables.

Table 1.7: Sensitivity analysis–varying CDS street lengths

	CDS street lengh			
	50m	100m	150m	200m
Fraction of the sample living in CDS neighborhood	0.073	0.107	0.124	0.131
Unrestricted model				
Marginal APE (eq. 1.3)	0.0973	0.0580	0.0321	0.0353
	(0.5190)	(0.1316)	(0.0592)	(0.0604)
Conditional APE (eq. 1.4)	0.0086	0.0065	0.0138	0.0107
	(0.2038)	(0.0652)	(0.0450)	(0.0417)
Correlation, ρ	-0.327	-0.255	-0.107	-0.146
Wald exogeneity test ($H_o : \rho = 0$)	0.062	0.275	0.115	0.223
	[0.804]	[0.600]	[0.735]	[0.637]
Restricted model ($\rho = 0$)				
APE	0.0118	0.0091	0.0150	0.0121
	(0.0079)	(0.0064)	(0.0065)**	(0.0061)**

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)
Standard errors are in parentheses; P-values are in square brackets.

First, note that the preferred length of 150m results in the lowest magnitude of the correlation coefficient, indicating that it is the best at capturing the unobserved determinants of neighborhood design and carpooling to school. Also note that, for all lengths, the Wald test fails to reject the exogeneity of D, implying that the model with ρ restricted to zero would produce more efficient estimates, which are shown on the last row. These estimates are all positive and similar in magnitude for all lengths, but only those resulting from CDS lengths of 150m and 200m are statistically significant. At first glance, this pattern is surprising since one would expect the effect to be stronger when the neighborhood boundary is closer to the CDS street end. However, recall from Table 1.2 that the estimation sample is highly unbalanced in the dependent variable. Shortening the neighborhood street length

shifts more observations to non-CDS status, further exacerbating the sample imbalance and reducing the variation in the treatment variable. This insight, hence, helps explain the lack of statistical significance for the shorter street lengths.

Omitted Variable Bias

The estimated effects may be confounded by potential omitted variables that factor into the negative correlation between the structural errors. The downward bias in the conditional APE relative to the marginal APE implies the omission of a variable that is correlated positively with CDS and negatively with carpooling. Two examples of such a variable are access to school bus service and access to public transit. If either of these services is more available in CDS areas, then it would also reduce the need for carpooling to school.

The CHTS does not contain data on school bus service. Instead, it is collected by the author.[5] This variable was not included in the main analysis because the school bus data reflect the year 2016, whereas the CHTS is from 2012, which was a year that experienced statewide cuts to school transportation budgets in California.[6] Assuming the budgets recovered with the economic recovery since then, the collected data my overstate the school bus availability in 2012

Second, access to public transit may have similar effects. Ideally, one would measure transit accessibility using information on routes, schedules, and congestion conditions using services such as Google Maps. However, use of online resources is restricted by NREL (the CHTS data vendor). Instead, the CHTS indicated which census tracts were within 0.25 and 0.5 miles from bus stops and rail stations, respectively. The caveat, however, is that this variable is not optimal as living close to a bus stop does not necessarily mean its routes serve the destination school.

[5]School bus data were collected from online sources such as the school and school district websites as well as www.greatschools.org.

[6]Los Angeles Times (2011, December 14). http://articles.latimes.com/2011/dec/14/local/la-me-california-budget-cuts-20111214. Accessed on April 3^{rd}, 2016

Table 1.8 compares the estimated effects from the main model specification (first column) to those from adding the school bus and public transit variables to it. Adding the variables does not change the general conclusions of the main analysis. An interesting observation is that adding the school bus variable drives the correlation coefficient closer to zero, implying that this variable further helps in capturing the unobserved factors determining the choices of neighborhood design and carpooling to school. In all specifications, the Wald test fails to reject the null hypothesis that CDS is exogenous. As such, the estimates under the restricted model are more efficient and slightly larger when the school bus variable is included. The findings from Table 1.8 do not necessarily rule out the potential for omitted variable bias, but if it exists, then it must be caused by variables other than school bus and transit accessibility.

Table 1.8: Sensitivity analysis–omitted variables

	Main specification	School bus	Transit	School bus + transit
Unrestricted model				
Marginal APE (eq. 1.3)	0.0321	0.0222	0.0346	0.0251
	(0.0592)	(0.0468)	(0.0655)	(0.0524)
Conditional APE (eq. 1.4)	0.0138	0.0164	0.0136	0.0160
	(0.0450)	(0.0427)	(0.0481)	(0.0456)
Correlation, ρ	-0.107	-0.038	-0.121	-0.057
Wald exogeneity test ($H_o : \rho = 0$)	0.115	0.016	0.128	0.031
	[0.735]	[0.901]	[0.720]	[0.860]
Restricted model ($\rho = 0$)				
APE	0.0150	0.0168	0.0150	0.0167
	(0.0065)**	(0.0068)**	(0.0054)**	(0.0067)**

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)
Standard errors are in parentheses; P-values are in square brackets.

Identification by Non-Linearity

As mentioned earlier, identification of the social capital equation (1.2) is possible theoretically by relying on the model's non-linearity. To ensure that the identification of the model stems from the instrumental variable and not the its non-linearity, a simple falsification test can be performed to compare the estimates with and without the instrument in the unrestricted model. If they are not affected, then it is a sign that the non-linearity is driving the model identification.

Table 1.9: Sensitivity analysis–non-linearity

	With instrument	Without instrument
CDS Coefficient, γ	0.4335	−0.3591
	(0.6021)	(1.0379)

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)
Standard errors are in parentheses; P-values are in square brackets.

As Table 1.9 shows, the coefficient estimates with and without the instrument have the opposite sign, indicating that the estimated coefficients and marginal effects in tables 1.5 and 1.6 are a result of the model being identified by the instrument and not its non-linearity.

1.5. Conclusion

Recognizing that social capital is a public good that can be prone inefficient private under provision, this paper tests suggestions that have been put forth in the urban economics and social capital literatures to use innovative urban designs to enhance social capital in communities. A simultaneous equations model is implemented to estimate the effect of neighborhood design, namely the cul-de-sac, on social capital, taking into account the endogenous sorting into neighborhood types. The paper offers two contribution: using carpooling as an objective measure of social capital that has not been previously used in the literature on the effects of the built environment on social capital and a precise spatial definition of the neighborhood through access to confidential geocoded household data from a California travel survey.

A quasi-replication exercise of a prior study in the literature is first performed using the proposed measure of social capital. The results of this exercise are in line with the replicated study as well as other studies in the literature, suggesting that carpooling can be considered a plausible measure of social capital.

The results of the main analysis yielded positive effects on carpooling to school that are consistent with the hypothesis that the certain neighborhood designs enhance social capital. A student living in a CDS neighborhood has a 1.5-percent higher probability of carpooling to

school than a comparable student living in a neighborhood of a different design. Considering that the probability of carpooling to school is 2.7% for the average student in the sample, the estimated effect is relatively substantial. On the other hand, the effect on carpooling to work is not statistically distinguishable from zero, which is not unreasonable since proximity and social interaction at the job location are expected to be the drivers of the decision to carpool as opposed to the residential social environment.

The analysis carries a few limitations. First, the instrumental variable (fraction of pre-1940 housing units in the county subdivision) that was utilized to address the endogenous self-selection into CDS neighborhoods hinges on a strong identifying assumption about the household's decision process for choosing a neighborhood and its design. It can also be argued that while this instrument can predict whether or not a neighborhood is a CDS, it may not necessarily predict the individual household's choice of neighborhood design. This distinction may undermine the instrument's relevance conceptually.

Second, statistical tests failed to reject the hypothesis that CDS is exogenous, owing to the rich set of control variables included in the model. Based on these tests, the statistically significant results were obtained from a model with the correlation between the errors restricted to zero and in which the instrument does not play an important role. These results effectively rely on statistical control for identification, which is a less than optimal identification strategy.

Third, the data lacks information on whether the carpooling is occurring exclusively between neighbors. Thus, the estimated effects might reflect factors influencing carpooling other than the neighborhood design and social environment.

The positive and statistically significant estimates of the effects held up consistently through the sensitivity analyses presented above, but these limitations present a challenge to attributing the results to neighborhood design. A causality claim may be made with more confidence if additional information is obtained on the social and spatial relations between

carpoolers or improvements are made to the identification strategy, such as employing more convincing instruments, using longitudinal data, or explicitly controlling for preferences and attitudes towards neighborhood design and social interaction.

Finally, the focus on the CDS design is motivated by its unique features, such as the perceived sense of security, safety, and familiarity, that may serve as the mechanism for enhancing the stock of social capital in the neighborhood. As such, the results cannot be generalized to other neighborhood designs. Broader insights can be gained by considering other neighborhood designs, other aspects of the built environment, or the presence of amenities that may promote social interaction.

This study is, nonetheless, an addition to a growing body of literature on the effects of the built environment on social capital. The findings suggest that the spatial configuration of residential communities (and not only spatial proximity between people) matters for enhancing social capital. The conclusion could potentially extend to any situation where the spatial configuration creates opportunities for social interaction for the people sharing that space, be it a neighborhood, public space, office space, or apartment floor. Therefore, the general policy implication is that social capital may be enhanced if urban planners, architects, and developers incorporated social interactions in their designs.

Chapter 2

Modeling Panel Count Data with Dynamics and Spatial Correlation

2.1. Introduction

The spatial econometrics literature has a well-established collection of models and estimators for analyzing continuous spatial data (see LeSage and Pace (2009) for a comprehensive textbook overview). While these models have been extended to some types of discrete data (e.g., binary and ordinal), there has been much less work on modeling spatial count data, which are common in studies of whether the variation between regions in the observed count of some event is a function of their location in space or proximity to one another. Count data also appear in situations where micro data on atomistic units of interest (e.g. persons, households, firms, etc.) are not available, so the researcher is forced to analyze aggregate data at the region-level (e.g. census tracts, zip codes, counties, etc.) instead where the counts represent the sum of individual observations of units located in each region.

This paper extends the Poisson-lognormal model to accommodate spatial correlation and temporal dynamics in panel count data and presents an efficient Bayesian Markov chain

Monte Carlo (MCMC) simulation algorithm to estimate its parameters. The dynamics can be in the form of separably additive time lags or the cumulative sum of lags, whichever is relevant for the application at hand. Basing the model on the Poisson process eliminates the need for linear transformations that alter the very nature of the observed count data. The spatial correlation is modeled through the covariance matrix of the random effects. The model also accommodates both within- and between-units overdispersion. After a series of simulation exercises, the model is applied to a dataset of zip codes in Southern California to study the spatial patterns of solar panel adoption. This application falls under studies of technology diffusion, and it has been the subject of recent research where linear transformations of the count data are common.

2.2. Related Literature

Early developments of spatial count models are found in the statistics literature. Data analysis is usually done in a hierarchical modeling framework where the count data follow a Poisson process whose conditional mean function includes a set of random effects assumed to follow a *conditional autoregressive (CAR)* procress to incorporate the spatial correlation (more discussion of CAR in the next section). Cressie (1993) and Banerjee et al. (2003) provide general overviews of this literature.

The spatial econometrics literature, on the other hand, has been experimenting with various approaches to modeling count data. Most commonly, empirical researchers often resort to transforming the observed count data into continuous data in order to take advantage of the well-established linear spatial models, their least-squares and likelihood-based estimators, and their specification tests. Others build upon the approaches from the statistics literature. Simões and Natário (2016) provide an overview of these approaches.

Other interesting approaches have also appeared in the spatial econometrics literature. Lambert et al. (2010) present a departure from the hierarchical modeling framework by

specifying a spatial lag model for the Poisson conditional mean directly, and they develop a two-step limited information maximum likelihood estimator. Similarly, Liesenfeld et al. (2016a) develop a panel count model with a spatially- and temporally-dependent Poisson latent variable. They estimate the model parameters using maximum likelihood, aided by an efficient importance sampling method that exploits the sparsity of spatial covariance matrices. Finally, Castro et al. (2012) adopt a different approach based on recasting the Poisson model into an ordinal probit model with a spatially-dependent latent variable and estimate it by maximizing the *composite marginal likelihood* function, which redefines the data as if they were observed in pairs.. The ordinal probit model is desirable because the model automatically accounts for overdispersion. The drawback, however, is the need to estimate the cut-points along the latent variable domain that define the observed ordinal categories. This feature presents a challenge in the context of count data: an assumption must be made about the value of upper-most cut-point above which the counts will be combined into a single category. Moreover, in cases where the observed counts are wide-ranging, further assumptions are needed regarding combining ranges of counts into representative categories in order to reduce the number of cut-points to be estimated.

2.3. Spatiotemporal Panel Count Model

2.3.1. Temporal Dependence

In addition to a static specification, the model developed here is capable of incorporating dynamics in two forms: separably additive lags and cumulative sum of lags.

Static Model

The Poisson distribution is the workhorse of models of count data, but its requirement that the conditional mean equal the conditional variance is at odds with the overdispersion often observed in most empirical applications. This overdispersion is interpreted as unobserved

heterogeneity in panel and spatial data. One way of accounting for it is by introducing unit-specific unobserved effects that are multiplicative in the conditional mean. For units $i = 1, 2, ..., N$ and time $t = 1, 2, ..., T$, the observed counts, y_{it}, are non-negative integers generated from the following model:

$$\begin{aligned} y_{it} \mid \mu_{it} &\sim Po(\mu_{it}) \\ \mu_{it} &= exp(x'_{it}\beta + a_i) = \alpha_i \, exp(x'_{it}\beta) \\ \alpha_i &= exp(a_i) \sim g(\alpha_i) \end{aligned} \tag{2.1}$$

where $g(.)$ is the assumed distribution of (the exponent of) the random effects, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_N)'$. The two most commonly assumed distributions of α_i are (i) gamma, resulting in the negative binomial model, and (ii) log-normal, resulting in the Poisson-lognormal mixture model. The latter is better suited to model spatial correlation. Thus, this paper will focus on the Poisson-lognormal mixture model.

The random effects model requires two key assumptions. First, conditional on the individual effect, α_i, all past, current, and future regressors, x_{it}, are *strictly exogenous*:

$$\mu_{it} \equiv E(y_{it}|x_{i1}, x_{i2}, ...x_{iT}, \alpha_i) = \alpha_i \, exp(x'_{it}\beta)$$

Second, the random effect, α_i, is assumed to be *uncorrelated* with the regressors, x_{it}. The joint probability function for unit i across all time periods is obtained by integrating out the random effect:

$$\begin{aligned} p(y_{i1}, y_{i2}, ..., y_{iT}) &= \int_0^\infty p(y_{i1}, y_{i2}, ..., y_{iT}, \alpha_i) \, d\alpha_i \\ &= \int_0^\infty p(y_{i1}, y_{i2}, ..., y_{iT} \mid \alpha_i) \, g(\alpha_i) \, d\alpha_i \\ &= \int_0^\infty \left[\prod_t p(y_{it} \mid \alpha_i) \right] g(\alpha_i) \, d\alpha_i \end{aligned} \tag{2.2}$$

Dynamics Through Separably Additive Lags

The strict-exogeneity assumption rules out the presence of temporally lagged counts in the conditional mean. To see this result, let one of the regressors be a lagged value of the outcome (i.e. $x_{it} = y_{i,t-1}$). The future value of the regressor is then $x_{i,t+1} = y_{i,t}$. Maintaining strict exogeneity implies that y_{it} is exogenous—an obvious violation. To accommodate the dynamics, a less restrictive *weak-exogeneity* assumption is made instead:

$$\mu_{it} \equiv E(y_{it}|x_{i1}, x_{i2}, ...x_{it}, \alpha_i) = \alpha_i \, exp(x'_{it}\beta)$$

where now only past and current values of the regressors, x_{it}, are assumed exogenous.

An important feature of the dynamic random effects model is the need to control for initial conditions, y_{i0}, as they are potentially correlated with the random effects. Wooldridge (2005) proposed conditioning the data and the random effects on the initial conditions as follows:

$$p(y_{i1}, y_{i2}, ..., y_{iT}, \alpha_i|X_i, y_{i0}) = p(y_{i1}, y_{i2}, ..., y_{iT}, |X_i, y_{i0}, \alpha_i) \, p(\alpha_i|X_i, y_{i0})$$

That way, the initial conditions y_{i0} are taken as given without the need to specify a distribution for them. There is, however, a need to specify the form of dependence between y_{i0} and α_i. One way of doing so is through a *conditionally correlated random effects* model (Mundlak, 1978), with the initial conditions added as a regressor:

$$\alpha_i = exp(\delta_0 y_{i0} + \bar{x}'_i\gamma + a_i)$$

where $\bar{x}_i$ denotes the average (over time T) of the time-varying covariates and a_i is a random effect that now represents the heterogeneity. Another benefit of Mundlak's correction is that it relaxes the second assumption above that the random effects are uncorrelated with the included regressors, which is difficult to justify in most applications. Stated differently, the correction assumes that the addition of the time-average regressors, which are sufficient

statistics for the covariates, controls for the potential correlation between the random effects and the regressors; therefore, any remaining unobserved effects can be assumed random and uncorrelated with the regressors (Cameron and Trivedi, 2013, p. 363).

Combining the exponents, the dynamic model can be written as:

$$\begin{aligned} y_{it} \mid \mu_{it} &\sim Po(\mu_{it}) \\ \mu_{it} &= exp(x'_{it}\beta + \delta y_{i,t-1} + \delta_0 y_{i0} + \bar{x}'_i\gamma + a_i), \end{aligned} \tag{2.3}$$

which is of the same form as the static model (2.1); therefore, the same techniques used for estimating it can be used for the dynamic model. A problem arises, however, from the presence of the lag in the exponent that may result in the model becoming explosive. Since the lag is non-negative, $\delta y_{i,t-1} \geq 0$ for $\delta > 0$. Moreover, if widely-varied counts across time periods are observed for unit i, the conditional mean function, μ_{it}, may exhibit sharp discontinuities that can result in a poor fit (Cameron and Trivedi, 2013, p. 370). This problem can be mitigated by replacing the lagged term with $\tilde{y}_{i,t-1} = \ln(y_{i,t-1}+1)$, where the addition of 1 is necessary for the case of $y_{i,t-1} = 0$.

Dynamics Through Cumulative Sum of Lags

In some contexts (as in the empirical application in this paper), it can be argued that the cumulative sum of past counts (denoted by B_{it}) influences the current count. With the *weak-exogeneity* assumption still maintained, the dynamic model can be adjusted as follows:

$$\begin{aligned} y_{it} \mid \mu_{it} &\sim Po(\mu_{it}) \\ \mu_{it} &= exp(x'_{it}\beta + \delta B_{it} + \delta_0 y_{i0} + \bar{x}'_i\gamma + a_i) \\ B_{it} &= \sum_{j=1}^{t-1} y_{ij} \end{aligned} \tag{2.4}$$

The same problem of explosive behavior arises as in the standard dynamic model. It can be mitigated similarly by replacing B_{it} with $\tilde{B}_{it} = \ln(B_{it} + 1)$.

2.3.2. Poisson-Lognormal Model

The standard (non-spatial) Poisson-lognormal model assumes a normal distribution for a_i, which induces a lognormal distribution, $\alpha_i = e^{a_i}$. Before proceeding, an important omission from the model (2.1) is worth addressing. As Cameron and Trivedi (2013, p.346) note, α is used in this model as a unit unobserved effect rather than an overdispersion parameter. In other words, it is used to account for the heterogeneity resulting from the panel structure of the data. While it may soak up the overdispersion across units, potential time-varying within-unit overdispersion may still persist. To account for it, a random-error term, ε_{it}, that varies across units and time may be included. The model, then, becomes:

$$\begin{aligned} y_{it} \mid \mu_{it} &\sim Po(\mu_{it}) \\ \mu_{it} &= exp(x'_{it}\beta + a_i + \varepsilon_{it}) \\ a &\sim \mathcal{N}_N(0, D) \\ \varepsilon_{it} &\sim \mathcal{N}(0, \sigma^2_\varepsilon) \end{aligned} \tag{2.5}$$

where x_{it} is a $(K \times 1)$ vector that contains an intercept term and the covariates (as well as the additional regressors associated with the dynamic models), and β is a vector of corresponding coefficients. D is a $(N \times N)$ covariance matrix that may be diagonal for a non-spatial model or unrestricted to account for spatial correlation.

The model can be equivalently written in *error-component* form, $\eta_{it} = a_i + \varepsilon_{it}$, with a combined covariance matrix, $\Sigma = D + \sigma^2_\varepsilon I_T$. However, Σ is a structured matrix, and estimating its components separately is more informative about the nature of correlation in the data.

Latent Variable Representation

For estimation purposes, obtaining the marginal likelihood analytically from (2.2) is not possible as the integral is intractable. Thus, frequentist estimation is usually done through

simulated likelihood or Gaussian quadrature methods (Cameron and Trivedi, 2013, p.122). Alternatively, Bayesian MCMC methods can be utilized to circumvent the evaluation of the integral and simulate the joint posterior distribution of the parameters. In this framework, the random effects are treated as parameters to be sampled to augment the data (Chib et al., 1998). However, the resulting full conditional posterior distribution of β is non-standard, requiring an Metropolis-Hastings (MH) step in the MCMC algorithm. An alternative approach that leads to a standard distribution is to adopt a latent-variable representation (Cameron and Trivedi, 2013, p.462):

$$\begin{aligned} y_{it} \mid z_{it} &\sim Po(exp(z_{it})) \\ z_{it} &\equiv ln(\mu_{it}) = x'_{it}\beta + a_i + \varepsilon_{it} \end{aligned} \tag{2.6}$$

where now the data augmentation procedure consists of sampling the latent variables, $\{z_i\}$, as well as the random effects, $\{a_i\}$.

Spatially-Correlated Random Effects

Spatial correlation is introduced into the model by specifying the following spatial autoregressive process for the random effects:

$$a = \rho W a + v \tag{2.7}$$

where ρ is the spatial correlation, W is the $(N \times N)$ spatial weight matrix defining the relationships between units, and v is a vector of errors. Model (2.11) implies that the random effect of one observational unit is a function of the random effects of other units related to it through W plus an error. If the unit of analysis is a region (e.g., zip code), then the elements of W can be based on contiguity, where two regions are considered neighbors if they share any part of a border:

$$w_{ik} = \begin{cases} 0 & \text{if} \quad i = k \\ 1 & \text{if} \quad i \sim k \text{ (i.e., } i \text{ and } k \text{ are neighbors)} \\ 0 & \text{if} \quad i \not\sim k \text{ (i.e., } i \text{ and } k \text{ are not neighbors)} \end{cases}$$

It is common practice to standardize W such that its rows sum to unity as it results in desirable properties that are computationally useful. The resulting row-standardized W also has an interpretive benefit in that the spatial lag variable (Wa) for unit i represents the average of random effects of its neighbors.

Define $A = I_N - \rho W$ and solve for the random effects to get

$$a = A^{-1}v \sim \mathcal{N}_N(0, D) \tag{2.8}$$

The nature of the spatial correlation is manifest in the covariance structure of D, with D_{ik} (the $[ik]^{th}$ element of D) representing the correlation between a_i and a_k. There are two approaches to specifying D. The first, introduced by Besag (1974), assumes the following *conditionally autoregressive (CAR)* prior distribution for the random effects explicitly:

$$a_i | \{a_j\}_{j \neq i} \sim \mathcal{N}\left(\rho \sum_{j \in N_i} w_{ij} a_j, \sigma_a^2\right)$$

From which the covariance matrix of the joint distribution (2.8) is derived to be $D_c = \sigma_a^2 A^{-1}$. This distribution then induces the following distribution for the errors: $v \sim \mathcal{N}_N\big(0, \sigma_a^2\, A\big)$, which are heteroscedastic and not independent. The second approach, called the *simultaneously autoregressive (SAR)*, works in reverse by assuming independently and identically distributed errors, namely $v \sim N_N(0, \sigma_a^2 I_N)$, and letting it induce a distribution for the random effects implicitly. The resulting covariance matrix is $D_s = \sigma_a^2 (A'A)^{-1}$. Note that without the spatial correlation (i.e., $\rho = 0$), both D_c and D_s become diagonal matrices and the model reduces to the non-spatial Poisson-lognormal model.

CAR vs. SAR

The CAR specification features more prominently in the statistics literature, which cites as one of its advantages the fact that, in a hierarchical model such as the one developed here, it lends itself readily to conditional sampling of the random effects in a Bayesian MCMC simulation. Furthermore, the errors (v) induced by CAR are necessarily heteroscedastic, which removes the researcher's discretion over assuming homoscedasticity. In addition, it can be shown that SAR can be recast as a CAR, but not vice-versa; thus, the latter is considered more general. The SAR specification, on the other hand, is favored by econometricians as it is analogous to the autoregressive disturbance process in time-series models where the serial correlation in the disturbances stems purely from the lag term while the shocks are maintained as random. Cressie (1993) and Banerjee et al. (2003) discuss the technical differences between the two approaches, and Wall (2004) provides an empirical illustration of the spatial structure implied by them.

In their discussion of the two specifications in a linear hierarchical model, Parent and LeSage (2008) state that choosing one or the other is "a matter of modeling preference, computational convenience, or empirical performance in any particular application." For their application, they estimate both specifications and choose the one favored by the data based on estimating the Bayes factor.

In this paper, the SAR specification will be adopted. Besides the appeal of the similarity to standard time-series models, the advantage cited above of the explicit conditional structure of the CAR approach will prove to be irrelevant for the model developed here. As will be seen in the next subsection, the proposed MCMC algorithm will rely on sampling the latent variables, $\{z_i\}$, marginally of the random effects, $\{a_i\}$. This approach allows for sampling the random effects jointly as a vector, not conditionally element-by-element. Finally, while not considered in this paper, heteroscedasticity can be easily incorporated into the SAR approach by simply specifying unit-specific variances, σ_{ai}^2.

In summary, the Poisson-lognormal model for panel data with spatially-correlated random effects can be written in matrix form (i.e. stacking over N and T) as:

$$\begin{aligned} y \mid Z &\sim Po(exp(Z)) \\ Z &= X\beta + Q\, a + \varepsilon \\ a &\sim \mathcal{N}_N(0, D) \\ \varepsilon &\sim \mathcal{N}_{NT}(0, \sigma_\varepsilon^2\, I_{NT}), \end{aligned} \tag{2.9}$$

where $Q = I_N \otimes 1_T$ with size $(NT \times N)$ and $D = \sigma_a^2 (A'A)^{-1}$. In the case of a dynamic model, the additional variables are simply part of X.

2.3.3. Estimation via Bayesian MCMC Simulation

Bayesian estimation proceeds by simulating draws from the joint posterior density of the model parameters with data augmentation. Let $\theta \equiv (\beta, \sigma_a^2, \sigma_\varepsilon^2, \rho)'$ denote the vector of parameters to be estimated. By Bayes' Theorem, the *augmented* joint posterior density of (z, a, θ) is proportional to the product of the *augmented* data likelihood function and their joint prior density:

$$\begin{aligned} p(Z, a, \theta|y) &\propto L(Z, a, \theta\,; y)\, p(Z, a, \theta) \\ &= L(Z, a, \theta\,; y)\, p(Z|a, \theta)\, p(a|\theta)\, p(\theta), \end{aligned} \tag{2.10}$$

For the density of the latent variables, $p(Z|a, \theta)$, it is desirable to marginalize out the random effect (a) to improve the mixing of Markov chains in the simulation (Chib and Carlin (1999) and Chib and Jeliazkov (2006)). To derive the marginalized distribution, $p(Z|\theta)$, let $\eta_i = (\eta_{i1}, \eta_{i2}, ..., \eta_{iT})'$ be the error-component form for unit i. The *within-unit* covariance matrix is:

$$\Omega_i \equiv E(\eta_i \eta_i' | x_i) = \begin{bmatrix} (D_{ii} + \sigma_\varepsilon^2) & D_{ii} & \dots & D_{ii} \\ D_{ii} & \ddots & & \vdots \\ \vdots & & \ddots & D_{ii} \\ D_{ii} & \dots & D_{ii} & (D_{ii} + \sigma_\varepsilon^2) \end{bmatrix}_{(T \times T)} = D_{ii} 1_T 1_T' + \sigma_\varepsilon^2 I_T,$$

where D_{ii} is the $[ii]^{th}$ element of D. For units i and k, let their *between-units* covariance matrix be $\Omega_{ik} \equiv E(\eta_i \eta_k' | X) = D_{ik} 1_T 1_T'$, which is of size $(T \times T)$ and captures the spatial correlation. Then, stacking all N units, the overall covariance matrix is:

$$\Omega = E(\eta \eta' | X) = \begin{bmatrix} \Omega_1 & \Omega_{12} & \dots & \Omega_{1N} \\ \Omega_{12} & \Omega_2 & & \vdots \\ \vdots & & \ddots & \Omega_{(N-1)N} \\ \Omega_{1N} & \dots & \Omega_{(N-1)N} & \Omega_N \end{bmatrix}_{(NT \times NT)} = (D \otimes 1_T 1_T') + \sigma_\varepsilon^2 \, I_{NT}$$

Finally, the distribution of Z, marginalized over the random effects, is:

$$Z = X\beta + \eta \sim \mathcal{N}_{NT}(X\beta, \Omega), \tag{2.11}$$

Note that for the non-spatial model, Ω becomes a block-diagonal matrix, with each block having $(\sigma_a^2 + \sigma_\varepsilon^2)$ on its diagonal and σ_a^2 as its off-diagonal elements.

While the latent variable representation makes it possible to sample the vector of random effects jointly, the drawback is that each vector z_i must be sampled *conditionally on all other* $\{z_j\}_{j \neq i}$ since they are correlated through Ω. To derive the conditional distribution of z_i, one can use standard properties of the multivariate Normal distribution. However, those properties are based on partitioning Ω and then repeatedly inverting large blocks from it of size $((NT - T) \times (NT - T))$, thus severely slowing down the algorithm. Instead, an

alternative, but equivalent, set of properties that are based on the precision matrix, $\Lambda \equiv \Omega^{-1}$, can be used (see Rue and Held (2005)). To see how, first re-arrange Z, $X\beta$, and Λ such that the i^{th} subvectors and submatrix are at the top:

$$\tilde{Z} = \begin{pmatrix} z_{i_{(T\times 1)}} \\ z_{j_{((NT-T)\times 1)}} \end{pmatrix}, \quad \tilde{X\beta} = \begin{pmatrix} (x_i\beta)_{(T\times 1)} \\ (x_j\beta)_{((NT-T)\times 1)} \end{pmatrix}, \quad \tilde{\Lambda} = \begin{bmatrix} \Lambda_{ii_{(T\times T)}} & \Lambda_{ij_{(T\times(NT-T))}} \\ \Lambda_{ji_{((NT-T)\times T)}} & \Lambda_{jj_{((NT-T)\times(NT-T))}} \end{bmatrix}$$

Then,

$$\begin{aligned} z_i|\{z_j\}_{j\neq i} \equiv z_{i.j} &\sim N_T(\bar{z}_i, \bar{\Omega}_i) \\ \bar{z}_i &= x_i\beta - \Lambda_{ii}^{-1}\Lambda_{ij}(z_j - x_j\beta) \\ \bar{\Omega}_i &= \Lambda_{ii}^{-1} \end{aligned} \tag{2.12}$$

Thus, Ω need only be inverted once to conditionally sample all $\{z_i\}$ in each MC cycle. Repeated inversion of Λ_{ii} of size $(T \times T)$ is still needed though not computationally intensive.

With the conditional distribution of z_i derived, the augmented joint posterior density (2.10), with the redundant variables and parameters suppressed from the conditioning sets, can be written as:

$$p(Z, a, \theta|y) \propto \left[\prod_i^N f_{Pois}(y_i|z_i)\ \phi_T(z_i|\bar{z}_i, \bar{\Omega}_i)\right] \phi_N(a|0, D)\ p(\beta)\ p(\sigma_a^2)\ p(\sigma_\varepsilon^2)\ p(\rho) \tag{2.13}$$

Note that for the non-spatial model, Λ_{ij} becomes a zero-matrix, eliminating the second term of $\bar{z}_i$ and thus no inversion would be needed. As a result, $\{z_i\}$ could be sampled independently (and, hence, more rapidly) of one another.

Priors

The prior for the spatial correlation parameter, ρ, can be specified as a uniform distribution over its feasible range, which is defined as the range of values that ensures the non-singularity of the matrix $A = (I_N - \rho W)$. Anselin (1988) showed that for a symmetric W, $\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})$, where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of

W. Furthermore, if W is row-standardized, then $\lambda_{max} = 1$. For a non-symmetric, row-standardized W (as is the case in this model), LeSage and Pace (2009) showed that the eigenvalues of W may be complex numbers, in which case $\rho \in (r_s^{-1}, 1)$, where r_s equals the smallest purely real eigenvalue.

Hence, the prior distributions assumed for the model parameters are:

$$\beta \sim N(\beta_0, B_0), \qquad \rho \sim U(r_s^{-1}, 1),$$
$$\sigma_a^2 \sim IG(c_1/2, d_1/2), \qquad \sigma_\varepsilon^2 \sim IG(c_2/2, d_2/2)$$

Full Conditional Posterior Distributions

The parameters $(\beta, \sigma_a^2, \sigma_\varepsilon^2)$ and the random effects, a, all have full conditional posterior distributions of standard forms. That is not the case for ρ and the latent variables, $\{z_i\}$; therefore, they must be sampled via MH steps embedded in the MCMC algorithm and using proposal densities that converge to the parameters' target densities. Using an independence-chain proposal density tailored to envelop the target density and centered at its mode results in an efficient algorithm. However, the mode must be estimated through maximization, which can be computationally intensive in for high-dimensional parameters such as $\{z_i\}$. A less efficient but faster algorithm generates candidate draws from a random-walk proposal density: $\theta^p = \theta^c + N(0, \tau)$, where θ^p is the proposed iterate, θ^c is the current iterate, and τ is a tuning parameter that shrinks or expands the spread of the standard-Normal deviate to achieve an acceptance rate between 0.25 and 0.5 (Chib and Greenberg, 1995). Derivation of the algorithm components can be found in the appendix.

MCMC Algorithm

The MCMC algorithm to simulate (2.15) proceeds as in the following algorithm, comprised of 3 (collapsed) sampling blocks:

1. Sample Z, β, and a in one (collapsed) block comprised of two sub-blocks:

   1.a) Jointly sample (Z, β) marginally of a as follows:

   - $\big[\{z_i\}|\{z_j\}_{j\neq i}, \beta, \sigma_a^2, \sigma_\varepsilon^2, y\big] \sim f_{Pois}(y_i|z_i)\ \phi_{\text{T}}(z_i|\bar{z}_i, \bar{\Omega}_i)$ via a MH-step for each $i = 1, 2, ..., N$
   - $\big[\beta|Z, \sigma_a^2, \sigma_\varepsilon^2, \rho\big] \sim N_{\text{K}}(\hat{\beta}, \hat{B})$, where

     $$\hat{B} = \big[B_0^{-1} + X'\Omega^{-1}X\big]^{-1} \quad \text{and} \quad \hat{\beta} = \hat{B}\big[B_0^{-1}\beta_0 + X'\Omega^{-1}Z\big]$$

   1.b) $\big[a|Z, \beta, \sigma_a^2, \sigma_\varepsilon^2, \rho, \big] \sim N(\hat{a}, \hat{V}_a)$

   where $\hat{V}_a = \big[D^{-1} + \frac{Q'Q}{\sigma_\varepsilon^2}\big]^{-1}$ and $\hat{a} = \frac{\hat{V}_a Q'(Z - X\beta)}{\sigma_\varepsilon^2}$

2. Jointly sample σ_a^2 and ρ conditionally on a:

   2.a) $[\sigma_a^2|\rho, a] \sim IG\big(\frac{1}{2}[c_1 + N]\ ,\ \frac{1}{2}[d_1 + a'A'Aa]\big)$

   2.b) $[\rho|\sigma_a^2, a] \sim \phi_N(Aa|0, \sigma_a^2 I_N)\ p_{\text{U}}(\rho)$ via a MH-step (If the proposed candidate for ρ falls outside the feasible region, discard it and re-draw until a feasible candidate is obtained).

3. $\sigma_\varepsilon^2|\beta, Z, a, \sim IG\big(\frac{1}{2}[c_2 + NT], \frac{1}{2}[d_2 + (Z - X\beta - Qa)'(Z - X\beta - Qa)]\big)$

For the non-spatial model, sub-block (2.b) simply drops out.

2.4. Simulation Exercise

To evaluate the performance of the algorithm, it is implemented using simulated data generated according to (2.9). The spatial and non-spatial models are fit, each with their

static and dynamic specifications. The spatial weight matrix, W, must be constructed somehow. Rather than generating artificial regions, a set of $N = 500$ actual zip codes from California are used, where the contiguity relationships are determined using a GIS algorithm. The average number of neighbors for each region comes out to be approximately 5.6 with a standard deviation of 1.9. Then, artificial count data are simulated using two covariates generated as follows:

$$x_{i1} = U(-2, 2) + N(0, 9)$$
$$x_{i2} = U(-2, 2) + N(0, 4)$$

The uniform means and large variances provide sufficient variation for the time-average covariates of the Mundlak correction. The time horizon, T, is set to 4 and the prior distributions are specified to be diffuse. The simulation results are presented in Table 2.1.

The results from the non-spatial model (ignoring the spatial correlation) are presented in the top panel of Table 2.1. The algorithm recovers the true values of the parameters of interest, namely $(\beta_1, \beta_2, \delta, \sigma_a^2, \sigma_\varepsilon^2)$, with accuracy. The intercept tends to be biased downward and less precise in the dynamic specifications. The coefficients of the initial condition and the Mundlak-correction variables $(\delta_0, \gamma_1, \gamma_2)$ are all biased downward but remain within two standard deviations. Finally, the spread of the proposal density for the latent variables had to be shrunk considerably in all models in order to achieve the desired acceptance rate.

Turning to the bottom panel, the algorithm performs as well for the spatial model in recovering the true values of the parameters. The estimate of the spatial correlation coefficient, ρ, shows a downward bias but still falls within one standard deviation of the true value. Similar to the non-spatial model results, $(\beta_1, \beta_2, \delta, \sigma_a^2, \sigma_\varepsilon^2)$ are recovered accurately and the other coefficients are biased downward but within two standard deviations. Also the spreads of the proposal densities for both the latent variables and the spatial correlation had to be shrunk considerably.

Table 2.1: Posterior mean and standard deviation estimates using simulated data

Parameter	True value	Static	One lag	Cumulative lags
Non-spatial model				
β_0	0.6	0.7425 (0.0492)	0.4150 (0.0756)	0.4034 (0.0793)
β_1	-0.5	-0.5036 (0.0093)	-0.5109 (0.0091)	-0.5095 (0.0084)
β_2	0.2	0.2050 (0.0100)	0.2039 (0.0115)	0.2049 (0.0111)
δ	0.3	-	0.3379 (0.0196)	0.3496 (0.0178)
δ_0	0.1	-	0.0782 (0.0470)	0.0915 (0.0455)
γ_1	-0.2	-	-0.1723 (0.0317)	-0.1382 (0.0306)
γ_2	0.4	-	0.3487 (0.0359)	0.3063 (0.0350)
σ_a^2	0.7	0.8342 (0.0690)	0.8089 (0.0730)	0.7663 (0.0716)
σ_ε^2	0.3	0.3077 (0.0231)	0.3466 (0.0229)	0.3370 (0.0202)
Tuning, τ_z, for $\{z_i\}$		0.08	0.08	0.07
Spatial model				
β_0	0.6	0.7386 (0.0756)	0.3984 (0.1001)	0.3903 (0.0995)
β_1	-0.5	-0.5000 (0.0088)	-0.5128 (0.0091)	-0.5096 (0.0085)
β_2	0.2	0.2054 (0.0108)	0.2063 (0.0111)	0.2055 (0.0110)
δ	0.3	-	0.3389 (0.0196)	0.3515 (0.0186)
δ_0	0.1	-	0.0816 (0.0437)	0.0981 (0.0427)
γ_1	-0.2	-	-0.1691 (0.0291)	-0.1317 (0.0302)
γ_2	0.4	-	0.3700 (0.0335)	0.3140 (0.0332)
σ_a^2	0.7	0.7191 (0.0644)	0.6766 (0.0674)	0.6406 (0.0647)
σ_ε^2	0.3	0.3039 (0.0229)	0.3480 (0.0222)	0.3400 (0.0211)
Spatial correlation, ρ	0.5	0.4443 (0.0677)	0.5041 (0.0643)	0.4694 (0.0664)
Tuning, τ_z, for $\{z_i\}$		0.08	0.08	0.07
Tuning, τ_ρ, for ρ		0.02	0.02	0.02

$N = 500; T = 4$; MC sample size = 10,000; burn-in = 2,000.

Implementation Issues

A few implementation issues are worth discussing. First, the time horizon was restricted to 4 periods for computational convenience. Increasing T to, say, 12 slowed down the algorithm severely. This limitation is mostly driven by the inversion of Ω and Λ_{ii}. Indirect methods of inversion such as using matrix decomposition may provide improvements. Second, the spatial correlation reduces the *effective* sample size. Ignoring the correlation in the non-spatial model should result in underestimating the posterior standard deviations. While this result can be seen by comparing the static models, it is not as clear in the dynamic models. In general, the differences in standard deviations are very small. The lack of appreciable differences may be an artifact of the simulated data, which needs further exploration.

2.5. Empirical Illustration: Solar Panel Adoption

In this section, the model and algorithm are applied to data from the California Solar Initiative (CSI). There has been an increased interest in understanding the patterns of solar panels adoption in recent years. In particular, some of the empirical literature is focused on whether neighborhood peer effects can explain some of the spatial clustering observed in the data. As a matter of policy, the presence of such effects would suggest a role for social interaction within neighborhoods in promoting sustainable energy sources more effectively.

2.5.1. Empirical Literature

The CSI data were featured in a study by Bollinger and Gillingham (2012) where they fit a linear panel model to estimate the effect of the *installed base* (i.e., the cumulative sum of lags) on the fraction of solar homes in a zip code. While they control for zip code-quarter fixed effects, they do not allow for correlation between zip codes. They find that, on average, an additional installation in a zip code has a small but positive effect on the probability of future adoption in that zip code. In an earlier working-paper version, Bollinger and Gillingham (2010) estimate a hazard model, and their findings are consistent with their latter conclusion. Richter (2013) fits a linear model (similar to Bollinger and Gillingham (2012)) to data from the UK and finds similar results.

Gillingham and Graziano (2015a) use data from Connecticut and a negative binomial model to estimate the effect of the installed base within a certain distance and time window on the number of new installations in a block group. They control for spatial correlation only partly by including a dummy variable for block-groups that are part of a wider-area promotion campaign. They find a positive relationship between adoption and the installed base that diminishes with distance and time. They infer that this is suggestive of neighborhood peer effects through social interaction and visibility.

Balta-Ozkan et al. (2015) employ a spatial econometrics approach (i.e. via a spatial weight matrix) to study solar panel adoption in the UK, but they log-transform the dependent count variable to estimate a linear model. Using cross-sectional data, they find statistically significant spillover effects in the adoption of solar panels between regions.

In contrast, the proposed model in this paper preserves the count nature of the data and incorporate the installed-base specification, all while also controlling for spatial correlation between regions that is often omitted in the literature.

2.5.2. CSI Background

California has actively promoted the shift to sustainable sources of energy in the last two decades. Given its geographic location and the abundance of sunshine that it enjoys, the state has dedicated considerable effort to harnessing solar energy. Much of this effort has focused on the demand side, implementing policies that aim to spur demand for solar panels through incentive programs.

The CSI is a cash rebate program that was launched in 2007 with a budget of over USD2 billion, aiming to achieve nearly 2 giga Watts (GW) of installed solar capacity. The initiative targeted customers of the state's three investor-owned utilities, offering rebates for existing homes as well as new and existing commercial, agricultural, government, and non-profit buildings. The incentive is structured by 10 levels such that early adopters receive a higher rebate, with the decline in incentive level triggered by achieving pre-set milestones based on the installed capacity.

2.5.3. Data

Data on installations are published on the CSI website, spanning the period from 2007 to 2016. Among other variables, the data include the zip code of the customer as well as the date of applying for the rebate, which is a proxy for the date of adoption. A panel is then constructed of the number of installations in zip codes by any preferred time unit.

The CSI data are supplemented by data from other sources: annual socioeconomic and built environment variables from the American Community Survey (ACS); monthly energy consumption by county from California Energy Commission; national average monthly electricity prices from the Energy Information Administration; 2012 presidential election results by county. Most of these variables are annual; therefore, the CSI installation counts are aggregated to annual counts.

The analysis is performed on the subset of zip codes of residential customers falling within the jurisdiction of the utility company Southern California Edison. After excluding zip codes with missing data and those with no neighbors, the final sample is comprised of 458 zip codes and 4 years (2011-2014; the ACS data are only available from 2011 and there were no new installations by SCE customers under the CSI program after 2014), resulting in 1,832 zip code-year observations.

2.5.4. Analysis Results

The model is estimated with the static and dynamic specifications. Year dummy variables are added to control for potential region-wide unobserved time effects that may influenced adoption (e.g. marketing campaigns or changes in incentive level). To avoid explosive behavior in the algorithm, the covariates that are not fractions are standardized to have a mean of zero and a standard deviation of one.

Non-Spatial Model

The results from the non-spatial model are presented in Table 2.2. Note that these are estimates of coefficients, not marginal effects. Therefore, meaningful discussion of effect size is not possible. However, the coefficient signs are interpretable.

Table 2.2: Posterior mean and standard deviation estimates from the non-spatial model

	Static	One lag	Cumulative lags
Intercept	-0.2155 (3.6743)	-0.8293 (3.6436)	-1.2529 (3.6388)
Fraction voted Obama 2012	0.8251 (1.1595)	0.2941 (0.5143)	0.4455 (0.4339)
Fraction owner-occupied	2.3448 (0.4572)	1.9014 (0.8509)	2.2058 (0.8595)
Fraction white	1.1229 (0.4586)	0.5136 (0.6152)	0.6041 (0.5798)
Fraction Black	-0.1211 (0.7122)	-0.9249 (1.5898)	-0.9465 (1.6292)
Fraction Asian	1.2416 (0.6291)	-1.3429 (1.5224)	-1.4718 (1.3921)
Fraction Hispanic	-1.5738 (0.4594)	-1.3897 (0.9491)	-1.9036 (0.9078)
Fraction with college degree	0.2351 (0.5770)	0.3219 (1.0193)	0.4086 (1.0456)
Unemployment rate	0.7429 (0.9156)	-0.0364 (1.1682)	0.3838 (1.1782)
Fraction of households with children under 6	-0.0197 (0.3521)	0.4766 (0.4404)	0.5765 (0.4290)
Number of housing units	0.8437 (0.0545)	0.3105 (0.5290)	0.1661 (0.5034)
Housing density per square-mile	-0.4324 (0.0784)	-0.9832 (1.0861)	-0.8622 (1.0548)
Median income ($\times$ USD 1000)	0.1909 (0.0825)	-0.2776 (0.1312)	-0.3106 (0.1308)
Median age	-0.3690 0.0786)	-0.1540 (0.1321)	-0.1710 (0.1344)
Average household size	0.0956 (0.1004)	0.3205 (0.1731)	0.3711 (0.1626)
Commute to work (minutes)	0.0025 (0.0467)	0.0478 (0.0719)	0.0308 (0.0731)
Electricity consumption	-0.0012 (0.1366)	1.1214 (0.5371)	1.1729 (0.5307)
Price of electricity	0.0286 (2.5920)	-0.0771 (2.6155)	-0.3893 (2.6201)
y_{t-1}	-	0.5852 (0.0529)	-
Installed base	-	-	0.8075 (0.0282)
y_0	-	0.0173 (0.0027)	0.0072 (0.0017)
σ_a^2	1.0372 (0.0837)	0.1376 (0.0319)	0.0855 (0.0099)
σ_ε^2	0.0670 (0.0052)	0.0965 (0.0100)	0.0926 (0.0070)
Tuning, τ_z, for $\{z_i\}$	0.04	0.04	0.04

$N = 458$ zip codes; $T = 4$ years; MC sample size = 10,000; burn-in = 2,000.

The installed-base coefficient is positive and statistically important, which is in line with the previous studies and suggestive of a neighborhood peer effect. The signs of the other coefficients that are of statistical import are generally consistent across the three specifications. Focusing on the installed-base specification (3rd column), zip codes with a higher fraction of owner-occupied units tend to have more installations, which is reasonable since homeowners have more power than renters over the decision to install. Higher fraction of hispanics is associated with fewer installations. Zip codes with higher median income have fewer installations, which is puzzling based on the presumption that installation is costly. However, the negative relationship may be explained by the fact that this cost is reduced by the rebate. Coupled with the expected savings on electricity bills, lower income households may be the ones more likely to install. The same rationale may explain the positive coefficient of the average household size. Higher electricity consumption is associated with more

installations, which is expected. Finally for the initial condition variable, a higher number of installations at the beginning of the sample period is associated with more installations later, which is reasonable.

Spatial Model

Next, the coefficient estimates from the spatial model are presented in Table 2.3. Across the three specifications, the spatial correlation estimate is positive and statistically different from zero, implying that zip codes are influenced by unobserved factors that their neighboring zip codes are subjected to. The presence of the correlation reduces the effective sample size, which should produce larger posterior standard deviations in the spatial model compared to the non-spatial model. The installed-base variable has a positive coefficient that is similar in magnitude to that from the non-spatial model. However, it has a larger posterior standard deviation as expected. The signs and statistical importance of the other covariates are similar to those from the non-spatial model with the difference in posterior standard deviations as well.

Discussion of Results

Dynamics

An interesting observation from the results is that the coefficients of the lagged-count (y_{t-1}) and the installed-base are both statistically important in their respective models, which raises an interesting question about the nature of the dynamics underlying the effect on adoption. On the one hand, the installed-base specification assumes that a household develops the intent and decision to install as a result of observing the accumulation of installations in its zip code over the period spanning the entire past. On the other hand, the lagged count specification suggest that the household is influenced more by *recent trends* in new installations. Both hypotheses are plausible but tell a different story about the mechanism of the social interaction effect and, thus, have different policy implications. If

Table 2.3: Posterior mean and standard deviation estimates from the spatial model

	Static	One lag	Cumulative lags
Intercept	0.2539 (3.7205)	-0.3731 (3.6311)	-0.8189 (3.6734)
Fraction voted Obama 2012	0.1873 (1.5739)	-0.1388 (0.8547)	0.0285 (0.6706)
Fraction owner-occupied	2.2403 (0.4872)	1.9288 (0.8933)	2.1820 (0.8796)
Fraction white	1.2106 (0.4685)	0.5620 (0.5808)	0.7446 (0.5819)
Fraction Black	-0.1630 (0.7927)	-0.6951 (1.7659)	-1.0995 (1.6571)
Fraction Asian	0.9652 (0.7026)	-1.1650 (1.5161)	-1.3662 (1.4549)
Fraction Hispanic	-1.9445 (0.4865)	-1.3347 (1.0001)	-1.8874 (0.8722)
Fraction with college degree	0.2802 (0.5960)	0.2542 (1.1452)	0.4004 (1.0197)
Unemployment rate	0.2192 (0.9168)	0.0113 (1.1515)	0.5049 (1.1108)
Fraction of households with children under 6	0.1405 (0.3677)	0.4772 (0.4608)	0.3513 (0.4345)
Number of housing units	0.8201 (0.0534)	0.3865 (0.5037)	0.2434 (0.5166)
Housing density per square-mile	-0.3628 (0.0859)	-1.2474 (1.0351)	-0.9718 (1.0812)
Median income ($\times$ USD 1000)	0.1554 (0.0901)	-0.2757 (0.1253)	-0.2841 (0.1212)
Median age	-0.3136 (0.0895)	-0.1669 (0.1480)	-0.1782 (0.1386)
Average household size	0.2064 (0.1109)	0.3170 (0.1741)	0.3465 (0.1682)
Commute to work (minutes)	-0.0584 (0.0501)	0.0457 (0.0712)	0.0455 (0.0648)
Electricity consumption	0.0998 (0.1770)	1.1550 (0.5530)	1.0951 (0.5165)
Price of electricity	-0.0055 (2.6122)	-0.1059 (2.5958)	-0.4908 (2.6345)
y_{t-1}	-	0.5632 (0.0453)	-
Installed base	-	-	0.7741 (0.0294)
y_0	-	0.0158 (0.0024)	0.0071 0(.0017)
Spatial correlation, ρ	0.4253 (0.0570)	0.5444 (0.0930)	0.5246 (0.0981)
σ_a^2	0.9112 (0.0796)	0.1199 (0.0238)	0.0771 (0.0101)
σ_ε^2	0.0670 0.0053)	0.0930 (0.0087)	0.0891 (0.0065)
Tuning, τ_z, for $\{z_i\}$	0.04	0.04	0.035
Tuning, τ_ρ, for ρ	0.015	0.012	0.012

$N = 458$ zip codes; $T = 4$ years; MC sample size = 10,000; burn-in = 2,000.

the installed-base hypothesis is supported more by the data, then a promotional program can exploit the influence of existing installations in a neighborhood regardless of how far in the past they have been in place. On the other hand, the lagged-count hypothesis suggests that such a program should discount the effect of installations that took place in the distant past. A specification test (e.g. via Bayesian model comparison) would be necessary in deciding which specification is more appropriate.

Identification

While the signs of the coefficients are not out of the ordinary and in line with other studies, this paper does not make any claims about the identifiability of the neighborhood peer effects. Generally, spatial models are plagued by three main identification challenges.

The first is the reflection problem highlighted by Manski (1993), which is a simultaneity problem in which the direction of the peer effect is not identifiable. This is not an issue in this paper by virtue of specifying the temporal dependence through past counts only.

The second challenge is the endogenous self-selection of peers based on, say, preferences towards residing close to other environmentally-conscious households. Including a rich set of covariates that proxy for environmental preferences may help control for any correlation between the error term and installed-base. Including the 2012 election data may capture some of the unobserved political orientation of zip code residents, but this is only a single variable that is also time-invariant, so by no means is it sufficient. Including a rich set of zip code fixed effects (as is done in other studies) is not possible as the random-effects model assumes away any fixed effects by construction through the assumption that the unobserved effects are uncorrelated with the included regressors. The model only partially relaxes this assumption by implementing the Mundlak correction.

The third challenge is the potential presence of correlated unobservables between households within a zip code (e.g. marketing campaigns targeted at neighborhoods rather than individual households). As already stated, including zip code fixed effects as a remedy is ruled out. Time dummy variables only control for region-wide effects. An additional remedy that is possible in principle is the inclusion of zip code-year dummy variables to control for any zip code-specific, time-varying effects, such as the targeted marketing campaigns. However, including those (1,374 of them) may result in overfitting.

The proposed model, nonetheless, goes a step further to address a related problem that has been overlooked in previous studies: correlated unobservables between the zip codes themselves. This correlation can still be the result of marketing campaigns targeted at neighborhoods in contiguous zip codes or be based on the economic conditions in the city or county. Rather than including higher (spatial) level dummy variables, this paper's approach models this correlation explicitly and allows for it to vary in intensity through the spatial lag variable defined by the spatial weight matrix.

2.6. Conclusion

A Poisson-lognormal model count data with spatially-correlated random effects is presented. The paper adds to the spatial econometrics toolkit a model that preserves the discrete nature of the data and incorporates time dynamics (both in the form of separately additive lags or cumulative sum of lags). The model parameters are estimated by an MCMC algorithm that exploits the hierarchical structure of the model, relying on the latent-variable representation and a collapsed-block sampling scheme. The algorithm is also capable of estimating the spatial and non-spatial components of the covariance matrix separately, enabling the researcher to gain more insight about the correlation structure in the data.

After a simulation exercise to evaluate the performance of the algorithm, the model is used to analyze the spatiotemporal patterns of solar panel adoption using data from Southern California. The results show the importance of controlling for the spatial correlation in the random effects in order to draw correct inferential conclusions about the neighborhood effect. The empirical results raise an interesting question about the proper specification for the dynamics—namely, lagged count vs. installed base—that has been overlooked in the literature. An immediate extension to this paper that can help resolve this question is to employ Bayesian model comparison methods (via estimation of the marginal likelihood and Bayes factor) to choose the specification that is supported by the data. Model comparison is also useful in testing the presence of spatial effects to warrant estimating a spatial model in the first place and, if so, in deciding between the SAR and CAR specifications.

The analyses reveals a few implementation issues that warrant further investigation. First, alternative matrix inversion techniques may potentially speed up the MCMC simulation to allow for panels with longer time horizons. Also, while the empirical results show that ignoring the spatial correlation leads to over-confidence in the estimates as expected, the results from the simulation exercise are not as clear and, thus, needs be explored further.

Chapter 3

Peer Effects and Spatial Correlation in Solar Panel Adoption

3.1. Introduction

Theoretical economic models of efficient-technology diffusion predict, with empirical support, a gradual diffusion process as consumers weigh the trade-off between the expected future benefits of the technology and the upfront cost of acquiring it (Jaffe and Stavins, 1994). When the diffusion process is slower than optimal, an *energy efficiency gap* forms between the optimal and realized levels of investment in the technology. The underlying behavior is that consumers undervalue the expected discounted future energy cost savings brought upon by the efficient technology, leading them to under-invest in it. This inefficient investment may be a result of a market failure in the form of imperfect information that biases the consumer's valuation.

Policies aimed at promoting adoption of efficient technologies are seen as addressing two market failures. First, adopting the efficient technology reduces the negative externalities, such as climate change and pollution, generated by the existing inefficient ones that it replaces

(albeit a second-best measure compared to a Pigouvian carbon tax). Second, consumer private welfare is improved by the energy cost savings. The question is how to address the information imperfection? As a general principle, a first-best intervention is one that tackles it directly via information disclosure. However, when it is ineffective in practice, subsidizing the efficient technology as a second-best can improve welfare (Allcott and Greenstone, 2012).

Solar photovoltaic (PV) panels, as a technology that generates clean renewable energy that directly replaces fossil fuel-based energy while at the same time immediately offsetting consumers' energy costs (and sometimes generating revenue for them), is a neat application of these concepts. The fact that it is a niche good that consumers are still in the process of accumulating their collective knowledge and experience about it, coupled with uncertainty about future fossil fuel energy prices and government energy policies, highlights the information imperfections associated with it and raises the stakes of the trade-off that consumers weigh in the face of the high upfront cost of equipment and installation. However, the recent exponential growth in solar panel adoption globally (Figure 1 in the appendix) has all the ingredients of the discussion above that make it a fertile ground for research. Interest lies in what, if any, may have contributed to addressing potential information imperfections in light of technological advances, financing solutions, and subsidies of various forms that coincided with the phenomenon and may have all contributed, to varying degrees, to lowering both direct (equipment and installation) and indirect (information) costs and spurred the demand.

Using data from the California Solar Initiative (CSI) rebate program in a spatial econometric framework to estimate a discrete-time transition model, this paper contributes to a small but growing literature on the role of peer effects play in the diffusion process. If peer effects are found to have played a meaningful role, controlling for other contributing factors, then the government and solar panel suppliers may consider leveraging consumers' social networks as a conduit for disseminating information and raising awareness about the benefits of solar energy, thus lowering information search and learning costs with an intervention

that is closer to the direct first-best. The sign of the peer effect estimated in this paper is consistent with that found in prior studies, but the effect is not statistically important. Further development of the estimation strategy—to address the outstanding identification issues as well as explicitly control for other determinants of adoption—is needed before a definitive conclusion can be stated about the existence and magnitude of the peer effect. However, the results do reveal that correlated unobservables, if left unaccounted for, lead to an upward bias in the peer effect estimate.

3.2. Related Literature

According to the International Renewable Energy Agency, global solar PV capacity grew from 14.6 to 385.7 giga Watts (GW) since 2008. The years 2015-2017 alone accounted for 58% of the 371 GW added, with net additions in 2017 more than doubling those in 2015 (IRENA, 2018). During the same period, solar PV module prices and installations costs have fallen substantially (Figure 2 and 3), thus reducing the direct portion of the upfront cost of adoption.

Various market solutions and government policies have also contributed to lowering this barrier. Most notably, third-party ownership (TPO)—a leasing arrangement whereby the solar contractor bears the installation cost and retains ownership of the solar system while the consumer reaps its benefit for a monthly payment—is a market innovation that is credited with attracting new consumer segments, especially younger, less educated, and lower income ones (Rai and Sigrin, 2013; Drury et al., 2012). Also, Ameli et al. (2017) and Kirkpatrick and Bennear (2014) find that the Property-Assessed Clean Energy program—loans provided by municipalities to homeowners to finance the purchase and installation of their solar systems and are repaid over 20 years through an increment added to their property tax—increases solar panel adoption. These findings imply that, for some consumers, TPO and financing arrangements solve a credit constraint problem that might have been part of slow diffusion

inefficiency. In addition, there have been policies designed to shape the incentive structure faced by energy users, such as cash subsidies, tax credits, tiered-electricity pricing, and net energy metering (NEM) that allows consumers to sell excess solar power back to the grid. Several studies find different combinations of these policies to be effective in accelerating adoption (Borenstein, 2017; Vaishnav et al., 2017; Matisoff and Johnson, 2017; Crago and Chernyakhovskiy, 2017).

There are also documented efforts by governments and suppliers to address the information imperfections. The TPO arrangement, for example, is seen as a way of mitigating the uncertainty associated with adopting a new technology by having the TPO assume the operational and maintenance duties on behalf of the host customer (Rai et al., 2016; Rai and Sigrin, 2013; Rai and Robinson, 2013). Information disclosure, in the form of marketing and information announcements, is also found to be effective in bridging the information gap (Reeves and Rai, 2018; Rai et al., 2016; Sigrin et al., 2015).

A subset of the studies cited above has uncovered, somewhat tangentially, suggestive evidence of peer effects on adoption decisions, which the literature that is exclusively focused on the subject appears to develop a consensus around. Studying adoption in California, Bollinger and Gillingham (2012) estimate the consumer's probability of adopting to be 0.78 percentage points higher in response to new installations in their zip code, reflecting a neighborhood peer effect. Using data from Connecticut, Gillingham and Graziano (2015b) find that the number of installations in a census block group increases by an average of 0.44 solar systems if an installation occurred in the prior year within half a mile. They show that the effect decays in space and time, attributing it to social interaction and the visibility of the solar panels. Richter (2013) finds a positive but very small effect on adoption in the United Kingdom. Noll et al. (2014) find positive peer effects on adoption as a result of leveraging trust networks in solar community organizations that are formed to encourage adoption by providing financing options and information about the benefits of solar PV's.

3.3. Data

3.3.1. California Solar Initiative

The CSI[1] was a statewide incentive program that provided cash rebates for new solar PV installations. Following Governor Arnold Schwarzenegger's Million Solar Roofs initiative, the state passed a law in 2006 allocating a budget of USD2,167 million to achieve 1,940 mega Watts (MW) of installed capacity by the year 2016, of which the vast majority (1,750 MW) were sought from the general market and the remaining 190 MW from two specialized programs for low-income residential customers.

The CSI targeted customers of California's 3 investor-owned utilities (Southern California Edison (SCE), Pacific Gas and Electric (PGE), and San Diego Gas and Electric (SDGE)) that represent 68% of the state's electric load. It covered installations in existing homes as well as new and existing commercial, industrial, government, non-profit, and agricultural properties, with the utilities administering the program directly under the oversight of the California Public Utilities Commission (CPUC). The CSI launched in 2007 and ran until the capacity or budget targets were reached for each utility.

Incentive Structure

The overall target for each utility is broken down into 10 intermediate sub-targets (called steps) of varying capacity to be achieved over the lifespan of the program. The incentive rate is lowered automatically each time a sub-target is reached. As such, the CSI program was designed to reward early adopters better than later ones. The cash rebate amount received by the system owner was determined based on the prevailing step incentive rate at the time of applying to the program as well as the physical attributes of the solar system. The targets, budgets, and incentive rates varied across utilities and sectors.

[1]The complete program details can be found in the CSI Program Handbook (2017)

Application Process

To encourage adoption, the CPUC designed a simplified application process. First, the customer conducts an online energy efficiency audit of their home to determine the need for and size of the solar system. Second, the customer hires a solar contractor from a searchable database of licensed contractors on the CSI website. Third, the customer applies for the incentive, often done by contractor on their behalf. Completing this step as early as possible is critical as it locks in the prevailing incentive rate on the date of the application. On this *reservation date*, the funds for the customer's rebate are reserved based on the size and estimated performance of the system. Fourth, the customer must install system within 12 months of the reservation date. Upon completing the installation, the utility interconnects the system to the grid, after which the customer begins to reap four types of benefits. First, the customer's electricity consumption is immediately offset by the solar energy they generate, lowering their electricity bill. Second, excess energy is fed back into the grid in exchange for credit applied towards the customer's electricity bill for further offsets. Third, and the final step in the application process, the customer claims the cash rebate if they own the system. Finally, they can claim federal or local tax credits they are eligible for.

3.3.2. Estimation Sample

The estimation is restricted to 7,379 SCE customers representing the entire population of the CSI general market participants from Orange County with accounts established before or during the first phase of the CSI. The data contain customers' street addresses as well as dates of installations, thus enabling the construction of a spatiotemporal estimation dataset. Table 3.1 presents the summary statistics of the main variables.

Table 3.1: Summary statistics ($N = 7,379$)

	Mean	(St. Dev.)
All electric appliances[a]	0.025	(0.156)
Pool at home[a]	0.338	(0.473)
Electric vehicle[a]	0.069	(0.253)
Energy efficiency program participation[a]	0.289	(0.453)
Energy efficiency total rebate (USD)	31.80	(201.91)
Reserved CSI incentive rate (USD per Watt)	0.948	(0.751)
CSI rebate (USD)	3869	(4035)
Install price (USD per Watt)	8.138	(2.956)
Total installation cost (thousand USD)	33.175	(16.143)
System size (kW)	4.600	(2.208)
Home owner[a]	0.988	(0.108)
Solar system owner[a]	0.554	(0.497)
Solar irradiation (MW per sq-meter)	2.071	(0.807)
Tax cap[a]	0.928	(0.258)
Tax credit if system owner (USD)	9621	(4361)

[a] Binary variables

Besides the CSI-related variables, the all-electric-appliances and pool variables aid in capturing irregularities in consumption. The electric vehicle and energy efficiency program participation variables serve as proxies for environmental preferences. Almost the entire sample is comprised of homeowners, reflecting the fact that renters seldom have the power to install solar panels over properties they don't own. Solar irradiation is a measure of the amount of energy from the sun that the customer receives at their location.

One noteworthy statistic is the fact that only 55% of customers own their CSI solar systems while the rest act as hosts to systems owned by TPO's. This fact has implications on how certain covariates are specified in the model as only system owners receive the CSI rebate and bear the upfront installation cost. Specifically, the incentive rate and install price must interacted with system ownership status. The tax credit also can only be claimed by the owner, which is already factored into calculating the credit value. In addition, the tax credit was capped at USD2,000 prior to the year 2009. Therefore, a dummy variable representing the tax credit cap must be interacted with the tax credit in order to capture changes in the effect before and after the cap was lifted.

Exposure vs. Reservation

The sample exhibits a few irregularities that may present challenges in interpreting the results. First, as mentioned earlier, the CSI incentive rate is a 10-step function that declines over time as more systems are installed. Second, not all customers are *exposed* to the CSI program from its launch to its conclusion. Third, customers *reserve* their rebates at different incentive rates at different points in time. Table 3.2 presents a breakdown of all 9,214 OC customers by exposure and reservation step.

Table 3.2: Exposure and reservations by step

Step	Incentive rate USD per Watt	Avg. installation price USD per Watt	Target MW	Step length Days	Exposed Percent	Reserved Percent
2	2.50	9.90	10.6	618	80.12	2.64
3	2.20	10.90	15.2	396	3.92	5.86
4	1.90	10.78	19.7	400	4.94	10.08
5	1.55	9.91	24.3	223	2.15	11.72
6	1.10	8.98	28.8	207	2.51	11.28
7	0.65	7.48	32.6	202	1.70	10.98
8	0.35	6.67	38.0	210	2.56	11.44
9	0.25	6.18	43.3	182	1.15	11.92
10	0.20	5.79	53.1	476	0.95	24.08

The vast majority (80%) of customers were exposed to the highest incentive rates available at the start of the CSI program. They include customers that were present prior to the CSI or those who moved to Orange County during step 2. The remaining 20% are customers who moved in during later steps, which are excluded from the main analysis. In contrast, the majority of customers reserved their rebates during the later steps of the program: about 58% customers installed during or after step 7, which went into effect in September of 2011, at incentive rates markedly lower than those during the first 3 steps that ended in July of 2010.

While the declining incentive rate over time was designed to encourage early adoptions, the observed influx of CSI reservations during the later steps (at low incentive rates) of the program may be driven by the dramatic decline in installation cost that was alluded to in Section 2. The second column in Table 3.2 presents the in-sample average installation price.

Throughout the CSI program, the average price dropped by 42%. Another factor that may explain the increased reservations during the later steps is that the MW targets for said steps were higher, requiring, on average, a larger number of customers to reach them.

3.3.3. Exploratory Duration Analysis

Before delving into the main analysis of peer effects, it is instructive to explore the data using standard continuous-time duration models, taking advantage of the availability of duration data in days. Specifically, these models shed light on the pattern of duration dependence as well as the effect of key time-varying covariates that cannot be modeled explicitly in the main analysis due to their lack of cross-sectional variation in a discrete-time setting.

Let T denote duration until *transition* from one state (no adoption) to another (adoption), and let t be its observed realization. The workhorse of these models is the hazard function, which is the instantaneous probability that T ends (i.e. transition occurs) at t conditional on having survived until t. Different distributional assumptions give rise to different shapes of the hazard function and, hence, the pattern of duration dependence.

Three models are used in this exploratory analysis. The first is the Weibull model, which is characterized by a monotonic hazard whose shape parameter determines if it is increasing (decreasing), indicating positive (negative) duration dependence. The second model assumes duration follows a lognormal distribution, leading to a hazard function that slopes upward initially then downward, thus relaxing the monotonicity assumption, albeit still a restrictive shape. The third is the piecewise constant hazard model that allows the hazard to flexibly take on different baseline values in different periods. In the three models, unobserved heterogeneity is incorporated via individual random effects assumed to follow a lognormal distribution with variance v.

Exploratory analysis results

It must be noted that since the time-varying covariates do not vary cross-sectionally, their effects are estimable under the Weibull and lognormal models only as they are absorbed by the step-specific constants in the piecewise constant hazard model. The maximum likelihood estimates from the three models of the effects of the covariates on expected log-duration are presented in Table 3.3. As such, the coefficients represent elasticities (for logarithmic covariates) and semi-elasticities (for linearly-specified covariates).

As mentioned earlier, the most valuable insights from this exploratory analysis are those related to time-varying covariates and duration dependence. A detailed discussion of all of the covariate effects is preserved for the main analysis of the peer effects model in Section 5.

The Weibull and lognormal models show that, for system owners, higher incentives and lower installed prices are associated with shorter duration—both are expected results. A Weibull shape parameter (α) that is greater (smaller) than 1 indicates a hazard that is monotonically increasing (decreasing). The lower panel of Table 3.3 reports an estimate of $\ln(\alpha)$ that is significantly greater than zero, indicating positive duration dependence. This pattern is corroborated by the estimates of the step-specific constants under the piecewise constant hazard the model, indicating an overall increasing baseline hazard with an intermediate flat region. Finally, the heterogeneity variance estimates are very small, with the likelihood ratio test failing to reject the null hypothesis on two of them. This result indicates that, after controlling for the observed covariates, the sample is fairly homogeneous.

Table 3.3: Exploratory analysis results ($N = 7,378$)

Model	Weibull		Lognormal		PWCH	
Outcome of interest	$E(\ln t\|x)$		$E(\ln t\|x)$		$E(\ln t\|x)$	
	Coef.	(SE)	Coef.	(SE)	Coef.	(SE)
Time-varying covariates						
Ln(incentive rate)	0.164	(0.034)***	1.270	(0.122)***	-	
Ln(incentive rate) × Owner	-0.113	(0.047)**	-0.397	(0.118)***	-	
Ln(install price)	-1.004	(0.19)***	-3.807	(0.393)***	-	
Ln(install price) × Owner	0.955	(0.178)***	2.187	(0.422)***	-	
County unemployment rate	0.023	(0.004)***	0.012	(0.006)**	-	
Time-invariant covariates						
All electric appliances[a]	-0.034	(0.016)**	-0.072	(0.034)**	-0.163	(0.080)**
Pool at home[a]	-0.017	(0.005)***	-0.025	(0.011)**	-0.107	(0.026)***
Electric vehicle[a]	-0.006	(0.009)	-0.017	(0.020)	0.003	(0.048)
Energy efficiency program[a]	0.045	(0.006)***	0.131	(0.012)***	0.168	(0.027)***
Log(tax credit)	-0.251	(0.014)***	-0.375	(0.023)***	-1.203	(0.055)***
No tax cap[a]	0.947	(0.027)***	1.682	(0.063)***	4.699	(0.091)***
Ln(tax credit) × No tax cap[a]	0.001	(0.003)	0.001	(0.008)	-0.067	(0.012)***
System owner[a]	0.333	(0.406)	-1.124	(0.914)	11.872	(0.517)***
Ln(system size)	0.195	(0.011)***	0.274	(0.015)***	1.086	(0.042)***
Ln(solar irradiation)	-0.001	(0.002)	-0.003	(0.005)	-0.008	(0.012)
Large contractor[a]	-0.048	(0.006)***	-0.115	(0.012)***	-0.300	(0.029)***
Zip code covariates (2012)						
Log(median income)	0.014	(0.025)	0.041	(0.051)	0.056	(0.127)
% with college degree	-0.001	(0.000)*	-0.001	(0.001)*	-0.003	(0.002)*
% Homeowners	0.001	(0.001)	0.001	(0.001)	0.004	(0.003)*
% Electric heating	0.000	(0.001)	0.001	(0.001)	0.001	(0.003)
% Single-family homes	-0.001	(0.001)	-0.001	(0.001)	-0.005	(0.003)
Intercept	8.325	(0.266)***	13.406	(0.810)***	-	
Ancillary parameters						
Weibull shape, $\ln(\alpha)$	1.605	(0.028)	-		-	
Lognormal scale, $\ln(\sigma)$	-		-1.083	(0.041)	-	
Step-specific constants, α_2	-		-		5.818	(0.489)***
α_3	-		-		3.389	(0.490)***
α_4	-		-		1.778	(0.496)***
α_5	-		-		0.779	(0.497)
α_6	-		-		0.527	(0.498)
α_7	-		-		0.288	(0.499)
α_8	-		-		-0.002	(0.500)
α_9	-		-		-0.495	(0.502)
α_{10}	-		-		-1.588	(0.503)***
Heterogeneity variance, v	0.000	(0.001)	0.037	(0.009)	0.018	(0.019)
LR test of $v = 0$ $[P - value]$	0.00	[1.000]	12.87	[0.002]***	0.99	[0.159]
log-likelihood	-54656.2		-55101.1		–54847.9	

Statistical significance: $*$ $(P < 0.10)$; $**$ $(P < 0.05)$; $***$ $(P < 0.01)$

[a] Binary variable.

3.4. Methodology

The spatial autoregressive process is the workhorse of the spatial econometrics framework. It can be assumed for the outcome variable, resulting in the spatial autoregressive (SAR) model:

$$y = \delta W y + x\beta + \varepsilon,$$

or the errors, resulting in the spatial error model (SEM):

$$\varepsilon = \rho W \varepsilon + v$$

where the spatial weight matrix, W, defines the relationships between individuals in the data. The terms (Wy) and $(W\varepsilon)$ are spatial lags, with their corresponding coefficients (δ and ρ) representing interaction effects between individuals or spatial correlation in the unobservable determinants of the outcome.

3.4.1. Spatial Transition Model with Peer Effects

The model implemented in this paper combines the SAR and SEM specifications into a model of transitions occurring at different points in time. Suppose the time horizon of the data is divided into discrete periods $q = 1, ..., Q$. In each period, the individual is faced with a binary choice of transitioning ($y_{iq} = 1$) or not ($y_{iq} = 0$). Consider the following latent-variable representing the difference in utility from the two choices:

$$\begin{aligned} z_q^0 &= \delta M_q y_q^1 + X_q^0 \gamma + \alpha_q \mathbb{1}_{N_q^0} + C_q a + \varepsilon_q^0 \\ \varepsilon_q^0 &= \rho W_q \varepsilon_q^0 + v_q^0 \end{aligned} \tag{3.1}$$

where the superscript "0" denotes individuals that have not yet transitioned at the beginning of interval q and "1" denotes those that have anytime prior to q. In each period, there are

N_q^0 of the first group and N_q^1 of the latter group. This formulation implies that those that have transitioned are discarded from the left-hand side in the next period, which is one way of representing the data as an unbalanced panel (i.e., the final period (Q_i) varies between individuals). Discarding transitioned individuals also rules out feedback effects of transitioning on time-varying covariates, thus maintaining their strict exogeneity assumption. In total, there are $N^0 = \sum_{q=1}^{Q} N_q^0$ observations used in the estimation and $N^1 = \sum_{q=1}^{Q} N_q^1$ cumulative transitions, all generated by N unique individuals.

W_q is the $(N_q^0 \times N_q^0)$ matrix of spatial relationships among individuals in interval q whereas M_q is the $(N_q^0 \times N_q^1)$ matrix of spatial relationships between them and their peers who have transitioned prior to q. Both matrices are constructed based on inverse distance. As such, nearby peers are more influential on an individual's choice to transition than those farther away. X_q^0 is a $(N_q^0 \times K)$ matrix of exogenous covariates and γ is their corresponding $(K \times 1)$ vector of coefficients; α_q is an period-specific constant; a is a vector of unobserved individual random effects and C_q is an incidence matrix that assigns them to their corresponding individuals in q; v_q^0 is a vector of normal *iid* errors with the variance normalized to an identity matrix. The model is a similar but extended version of the one proposed by Elhorst et al. (2017).

The model can be written in stacked (over q) format as follows (with the superscripts suppressed):

$$\begin{aligned} y &= \mathbb{1}\{z > 0\} \\ z &= \mathbb{X}\beta + Ca + \varepsilon \\ a &\sim \mathcal{N}(0, \sigma_a^2 I_N) \\ \varepsilon &\sim \mathcal{N}(0, (A'A)^{-1}) \end{aligned} \tag{3.2}$$

where

$$\mathbb{X} = \begin{bmatrix} M\mathbb{1}_{N^1} & X & C_\alpha \end{bmatrix}$$
$$\beta = \begin{bmatrix} \delta & \gamma' & \alpha' \end{bmatrix}'$$
$$A = I_{N^0} - \rho W$$

The vectors z and y are of size $(N^0 \times 1)$, W is a $(N^0 \times N^0)$ block-diagonal matrix with W_q occupying each block, and M is a rectangular matrix of size $(N^0 \times N^1)$ with M_q on its diagonal blocks. The combined coefficient vector β has $\tilde{K} = 1 + K + Q$ elements, including the $(Q \times 1)$ vector of period-specific constants, α. The $(N^0 \times \tilde{K})$ matrix $\mathbb{X}$ combines all of the regressors, including the $(N^0 \times Q)$ incidence matrix C_α that assigns appropriate period-specific constant. Similarly, C is a $(N^0 \times N)$ incidence matrix that assigns individuals their corresponding random effects. A is a non-singular matrix that induces the spatial correlation in the reduced-form errors.

The quantity of interest is the conditional mean of the outcome y, which is also the probability that transition occurs in period q conditional on surviving up until interval $q-1$:

$$E(y_q|\mathbb{X}_q, a, \beta) = \Pr(t_{q-1} \leq T < t_q | T \geq t_{q-1}, x) = \Phi(\mathbb{X}_q\beta + C_q a) \tag{3.3}$$

where $\Phi(.)$ is the standard normal cdf, resulting in a binary choice probit model of unbalanced panel data. Note that (3.3) is analogous to the hazard function in standard duration models discussed in the previous section.

Identification Issues

Spatial models of peer effects suffer from three identification challenges: reflection, correlated unobservables, and self-selection. Reflection is when peers influence one another simultaneously such that the direction of the effect is not discernible (Manski, 1993). Correlated unobservables is a situation whereby peers are subjected to common time-varying shocks such as a marketing campaign targeting their area. Self-selection is a result of en-

dogenous group formation whereby peers choose to reside near each other based on, say, shared environmental consciousness. Bollinger and Gillingham (2012) provide a mathematical illustration of these challenges and their threat to the consistency of the peer effect estimate.

In this paper, the reflection problem is addressed by discarding transitioned individuals from the left-hand side in each period; hence, only past transitions can influence individuals remaining in the sample, thus eliminating the simulteneity. As such, past transitions are effectively temporally exogenous with respect to the current period. Correlated unobservables are accounted for through the spatial autoregressive process of the errors. Self-selection, though, remains a threat to identification; thus, the estimated peer effect may possibly be confounded by omitted factors that may be driving the transition decisions of peers, such as shared preferences.

3.4.2. Estimation

Due to the spatial correlation in the errors, the likelihood function resulting from the model is not a simple product of independent likelihood contributions. As a result, integrating out the latent variable involves evaluating a N-dimensional integral, rendering likelihood maximization prohibitively infeasible. Instead, researchers usually resort to simulated-likelihood methods (Elhorst et al., 2017; Liesenfeld et al., 2016b; Franzese et al., 2016).

Alternatively, likelihood evaluation can be avoided altogether in a Bayesian framework whereby the joint posterior distribution of the parameters, given the data, is simulated via Markov-chain Monte Carlo (MCMC) methods (LeSage and Pace, 2009). Rather than integrating out the latent variable and random effects, they are estimated alongside the model parameters. The procedure starts with deriving the *augmented* joint posterior density using Bayes' theorem:

$$\begin{aligned} p(z, a, \theta|y) &\propto L(z, a, \theta; y)p(z, a, \theta) \\ &= L(z, a, \theta; y)p(z|a, \theta)p(a|\theta)p(\theta) \end{aligned} \tag{3.4}$$

where $\theta \equiv (\beta, \rho, \sigma_a^2)$. Since y is determined solely by z, the likelihood function can be written as $f(y|z)$ where $f(.)$ is the appropriate probability density function (pdf). Also, the latent variables can be marginalized over the random effects as follows:

$$\begin{aligned} z &= \mathbb{X}\beta + \eta \\ \eta &\sim \mathcal{N}_{N^0}(0, \Omega) \\ \Omega &= \sigma_a^2 CC' + (A'A)^{-1} \end{aligned} \tag{3.5}$$

which leads to efficiency improvements during the mixing of the Markov chains (Chib and Carlin, 1999; Chib and Jeliazkov, 2006).

The parameters are assumed to be independent with a joint prior density $p(\theta)$. The prior distribution of ρ is specified as uniform over the region that ensures non-singularity of A. For the non-symmetric, row-standardized W used in this model, LeSage and Pace (2009) showed that $\rho \in (r^{-1}, 1)$ where r is the smallest purely real eigenvalue of W. The prior distributions of β and σ_a^2 are assumed to be $\mathcal{N}_{\tilde{K}}(\beta_0, B_0)$ and $IG(c/2, d/2)$, respectively.

The joint posterior density then takes the following final form:

$$p(z, v, \theta|y) \propto \left[\mathbb{1}\{z > 0\}\right] \phi_{N^0}(z|\mathbb{X}\beta, \Omega) \phi_N(a|0, \sigma_a^2 I_N) \phi_{\tilde{K}}(\beta|\beta_0, B_0) f_{IG}(\sigma_a^2|c/2, d/2) f_U(\rho|r^{-1}, 1) \tag{3.6}$$

MCMC Simulation Algorithm

The joint posterior density (3.6) is simulated using the following collapsed block-sampling algorithm:

1. Jointly sample (z, β) marginally of a in one (collapsed) block:
   - $\big[z|\beta, \sigma_a^2, \rho, y\big] \sim \mathcal{TN}_{\mathcal{R}}(\mathbb{X}\beta, \Omega)$ where the elements $\{z_{iq}\}$ are sampled conditionally on one another from the appropriate region of truncation, R.
   - $\big[\beta|z, \sigma_a^2, \rho\big] \sim N_{\bar{K}}(\hat{\beta}, \hat{B})$, where $\hat{B} = \big[B_0^{-1} + \mathbb{X}'\Omega^{-1}\mathbb{X}\big]^{-1}$ and $\hat{\beta} = \hat{B}\big[B_0^{-1}\beta_0 + \mathbb{X}'\Omega^{-1}z\big]$.
2. $\big[a|z, \beta, \sigma_a^2, \rho\big] \sim \mathcal{N}_N(\hat{a}, V_a)$, where $V_a = \big[\frac{I_N}{\sigma_a^2} + C'A'AC\big]^{-1}$ and $\hat{a} = V_a C'A'A(z - \mathbb{X}\beta)$
3. $[\sigma_a^2|a] \sim IG\big(\frac{c+N}{2}\ ,\ \frac{d+a'a}{2}\big)$
4. $\big[\rho|z, \beta, a\big] \sim \phi_{N^0}\big(z - \mathbb{X}\beta - Ca|0, (A'A)^{-1}\big)\ p_{\mathrm{U}}(\rho) \equiv f(\rho|.)$ via a Metropolis-Hastings step (If the proposed candidate for ρ falls outside its feasible domain, discard it and re-draw until a feasible candidate is obtained).

The algorithm is started at some initial values for the parameters, and then cycled repeatedly to obtain $b = 1, ..., B_{MC}$ draws of each of the parameters making up their joint posterior distribution. The posterior means and standard deviations of the parameters can be computed after discarding earlier draws as burn-in while the chains stabilize and converge.

Marginal Effects

Coefficient estimates from nonlinear models are only informative about the existence and direction of the effects of their respective covariates. Their magnitudes are obtained by calculating the marginal effect on individual i's transition probability in period q from a change in each covariate x_{iqk}, which equals the partial derivative of equation (3.3),

$$\frac{\partial E(y_{iq}|x_{iq}, a_i, \beta)}{\partial x_{iqk}} = \phi(x_{iq}'\beta + a_i)\beta_k \tag{3.7}$$

This marginal effect is then averaged over all individuals and periods to obtain the average marginal effect (AME). This formula can be embedded in the MCMC algorithm where the AME is evaluated at each MC pass and stored. The posterior means and standard deviations of AME's are then obtained after discarding the initial burn-in values.

3.5. Estimation Results

Three model specifications are employed: (i) a standard non-spatial probit model ($\delta = \rho = 0$), (ii) a model of peer effects without spatial correlation ($\rho = 0$), and (iii) the full model ($\delta \neq 0$ and $\rho \neq 0$). The elements of W and M equal the inverse of the euclidean distance between individuals up to 2 kilometers and zero for those residing farther apart. The estimation sample is restricted to the years from 2009 to 2013 because (a) the final year of the CSI (2014) consists of transitions only, thus predicting the outcome perfectly, and, similarly, (b) the first two years (2007-2008) contain only a few transitions that may cause stability problems in the simulation. Nevertheless, while these transitions are discarded from the left-hand side, they are retained on the right-hand side as initial conditions making up the spatial lag ($M_1 y_1^1$) in the first period of the analysis (2009).

Due to high computational costs arising from the need to invert the large covariance matrix Ω, the estimation is performed on only a subset of the sample (region 5 in Figure C.4 in Appendix C). In addition, all of the covariates are time-invariant except the spatial lag. The time-varying ones can not be modeled explicitly since they do not vary cross-sectionally, and, thus, they are perfectly collinear with the year-specific constants. The posterior AME estimates are presented in Table 3.4.

Focusing on the preferred specification (iii), the neighborhood peer effect is positive but not statistically important. The year-specific constant increase over time, showing a positive duration dependence pattern, which is consistent with the findings of the exploratory duration analysis of Section 3.3. The spatial correlation (ρ) is estimated to be 0.378 and

Table 3.4: Posterior average marginal effects

Models	(i)		(ii)		(iii)	
Posterior estimates	AME	(SD)	AME	(SD)	AME	(SD)
Neighborhood peer effect	-		0.620	(0.445)	0.456	(0.475)
All electric appliances[a]	0.046	(0.062)	0.050	(0.063)	0.047	(0.062)
Pool at home[a]	0.016	(0.012)	0.014	(0.012)	0.013	(0.012)
Pool at home missing	0.087	(0.022)	0.087	(0.023)	0.082	(0.027)
Electric vehicle[a]	-0.035	(0.022)	-0.035	(0.022)	-0.035	(0.023)
Energy efficiency enrollment[a]	-0.055	(0.012)	-0.056	(0.012)	-0.055	(0.012)
Log(tax credit)	0.268	(0.020)	0.271	(0.021)	0.268	(0.021)
Log(system size)	-0.237	(0.015)	-0.237	(0.015)	-0.237	(0.016)
System owner[a]	-2.537	(0.185)	-2.567	(0.189)	-2.542	(0.195)
Log(solar irradiation)	0.010	(0.007)	0.010	(0.007)	0.009	(0.007)
Large contractor[a]	0.051	(0.012)	0.051	(0.012)	0.050	(0.012)
Zip code variables (2012)						
Log(median income)	0.491	(0.153)	0.510	(0.153)	0.493	(0.177)
% with college degree	-0.004	(0.002)	-0.004	(0.002)	-0.004	(0.002)
% Homeowners	-0.012	(0.003)	-0.012	(0.003)	-0.012	(0.003)
% Electric heating	0.002	(0.002)	0.002	(0.002)	0.002	(0.003)
% Single-family homes	0.007	(0.059)	0.007	(0.002)	0.007	(0.003)
Year-specific constants, α_q						
$q = 2009$	-1.863	(0.531)	-1.925	(0.534)	-1.835	(0.621)
$q = 2010$	-1.666	(0.531)	-1.728	(0.534)	-1.637	(0.621)
$q = 2011$	-1.565	(0.530)	-1.627	(0.534)	-1.535	(0.621)
$q = 2012$	-1.428	(0.530)	-1.492	(0.535)	-1.401	(0.621)
$q = 2013$	-1.091	(0.530)	-1.158	(0.534)	-1.064	(0.621)
δ	-		2.982	(2.144)	2.211	(2.301)
ρ	-		-		0.379	(0.069)
σ_a^2	0.239	(0.059)	0.248	(0.067)	0.249	(0.068)

Statistical significance: $*$ $(P < 0.10)$; $**$ $(P < 0.05)$; $***$ $(P < 0.01)$
$N^0 = 5,399$; $N = 1,494$; $B_{MC} = 10,000$; burn-in = 2,000.
[a] Binary variable.

statistically different from zero, providing evidence of correlated unobservables playing a role in individuals' decisions to install.

Turning to the time-invariant covariates, the effects of having only electric appliances or a pool (intended to capture abnormal demand for energy) are both positive, as expected, but not statistically important. Having an electric vehicle or enrollment in energy efficiency programs both have negative effects, but only the latter is statistically important. The negative sign is runs counter to the intuition behind including them as proxies for preferences of environmentally conscious individuals towards energy efficiency. One plausible explanation is that enrollment in efficiency programs lowers the need to install solar panels. In other words, the marginal benefit, in terms of efficiency gains and cost savings, of adopting solar energy

is unappealingly diminished for already efficient homes. In contrast, an inefficient home consumes more energy and, thus, faces a higher average electricity price than a comparable efficient home, ceteris paribus, because of California's tiered electricity pricing system. As a result, the inefficient home has a higher incentive to adopt solar energy because it replaces otherwise high-cost energy.

A 10% increase in the tax credit is associated with a 2.68-percentage point higher probability of adoption. Homes determined to need larger solar systems are less likely to adopt, which is expected from a cost perspective. System ownership shows the largest effect; the adoption probability is 254% lower for system owners compared to those in TPO arrangements, also reflecting the upfront cost barrier that owners have to overcome. To complement this finding, customers who hire a large contractor have a 5% higher adoption probability, reflecting the market power of these contractors enabling them to compete on prices as well as offer discounts and financing facilities.

The effects of zip code variables tell conflicting, if not puzzling, stories. Customers from high median income zip codes are more likely to adopt, which is reasonable from an affordability perspective, but runs counter to the argument that the incentive rebates, tax credits, and TPO arrangement attract low-income customers. In contrast, zip codes with a larger fraction of college graduates are associated with lower adoption probability, albeit the effect is very small. Zip codes with higher fraction of single-family homes are associated with higher adoption probability, which is expected. Finally, the most puzzling result is the negative effect corresponding to zip codes with higher fraction of homeowners. Since, compared to renters, homeowners have the power over the decision to install solar panels, a positive effect would be expected instead.

Discussion of Spatial Effects

To examine the role the peer effect plays in the results, first compare the estimates from the non-spatial model (i) to those from the peer effects model (ii). The effects of the

time-invariant covariates are statistically indistinguishable, but the year-specific constant are slightly smaller under model (ii). This difference suggests that, if not modeled explicitly, the peer effect is absorbed by the year-specific constants. However, when spatial correlation in the errors is accounted for in model (iii), the year-specific constants return to their levels under (i) but become less precise, which is expected as the correlation reduces the effective sample size. A more notable observation is the smaller magnitude of the peer effect under (iii) compared to (ii), suggesting an upward bias when the spatial correlation in the errors is not controlled for. This bias is evidence of the presence of correlated unobservables that account for part of the peers' decisions to adopt. As an example, a contractor might target an area with a marketing campaign. If it caused peers to adopt, then failure to control for the campaign or, if unobservable, the fact that spatially-proximate customers were targeted by the same campaign, then the estimation would overstate the peer effect.

Simulation Issues

The analysis could be enriched further by using a larger sample to possibly provide more spatial variation. Moreover, expanding the geographic coverage of the sample to include customers from different utility jurisdictions would enable explicit modeling of important time-varying factors such as the incentive rate and installed price that are otherwise absorbed in the year-specific constants in the current analysis. However, as alluded to earlier, expanding the sample increases the computational demands of the simulation rapidly, owing to the need to invert and store the large covariance matrix, Ω, in each MCMC pass.

The computation of the determinant of A required for sampling ρ is also computationally intensive. However, since its feasible domain is finite ($\rho \in (r^{-1}, 1)$), this burden is alleviated in the analysis above by calculating the determinant for a grid of ρ spanning its domain outside the MCMC algorithm. Then for each proposed candidate ρ in the MH-step, the determinant is obtained by interpolating the pre-calculated determinant values corresponding to the two closest ρ values in the grid (Pace and Barry, 1997). The gains in computation

speed justify the loss of precision resulting from the interpolation, which can be mitigated by using a sufficiently fine grid of ρ.

Finally, the stability of the Markov chains proved very sensitive to the assumed hyper-parameters of the inverse-gamma prior for the unobserved random effect variance, σ_a^2. Specifically, the chains exhibit explosive behavior with outlying draws from the right tail of the posterior inverse-gamma density. This behavior is suggestive of a true variance that is very small or one whose distribution is highly concentrated around its mode. In order to gain more insight into the issue, the non-spatial model (i) was estimated via maximum likelihood using the 5 subsamples representing the regions of Orange County shown in Figure 5. The estimates of σ_a^2 ranged between 0.16 and 0.67, which are relatively small. This finding implies that, conditional on the included variables, these subsamples are relatively homogeneous, which is consistent with the results from the exploratory analysis in Section 3 and the fact that the sample only represents solar customers who are potentially endogenously sorted based on shared preferences and attitudes towards solar energy. To facilitate the simulation, the hyper-parameters were chosen to be $c = 40$ and $d = 8$, resulting in a highly concentrated prior distribution with a mean of 0.21 and variance of 0.0025. Essentially, this is a very informative prior based on knowledge about the heterogeneity gleaned from subsamples from the other regions. More importantly, this discussion reveals that, for the data at hand, the inverse-gamma may not be a suitable prior distribution. A more agnostic analysis would employ a less concentrated prior distribution that is also more accommodating of small variances (potentially at the expense of conjugacy or parsimony). Gelman (2006) discusses possible alternatives.

3.6. Conclusion

The literature on the diffusion of new, energy-efficient technologies points to a market failure whereby imperfectly-informed consumers undervalue the expected future benefits of

the technology, leading to inefficient investment in it and, hence, a slower-than-optimal diffusion process. Second-best policies based on subsidizing adoption of the technology are common if the first-best, in the form of information provision, proves to be practically ineffective. A growing literature finds a role for peer effects in consumers' decision process of adoption that may be effective in addressing the information imperfection. If so, governments and suppliers could leverage consumers' social networks as an information channel. This paper contributes to this literature by estimating a transition model of neighborhood peer effects, controlling for spatial correlation, using data from the California Solar Initiative rebate program and Bayesian MCMC simulation methods. The results show a positive but not statistically important peer effect. However, they also reveal that spatially correlated unobservables bias the peer effect upward if not controlled for.

Further development of the model and analysis are necessary in order to draw definitive conclusions about the existence and magnitude of the peer effect. Most importantly, the model as it stands lacks clear identification, resulting in the estimates likely being confounded by self-selection effects resulting from the endogenous formation of peer groups. Second, the sample can be expanded to cover a wider geographic area to possibly provide variation in the unobserved random effects that would give insights about the role heterogeneity plays in the rate of adoption. Finally, data from utilities other than SCE would introduce cross-sectional variation in key time-varying covariates, such as the incentive rate and prices. While the effects of these covariates are accounted for in the current analysis through the year-specific constants, the ability to model them explicitly would facilitate making comprehensive conclusions about how much each of the cash rebate, prices, and peer effects contribute to the adoption process. These outstanding issues are to be undertaken in future extension of this research project.

Bibliography

Allcott, H. and Greenstone, M. (2012). Is there an energy efficiency gap? *Journal of Economic Perspectives*, 26(1):3–28.

Ameli, N., Pisu, M., and Kammen, D. M. (2017). Can the us keep the pace? a natural experiment in accelerating the growth of solar electricity. *Applied Energy*, 191:163–169.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Kluwer Academic Publishers, Dordrecht, The Nedherlands.

Balta-Ozkan, N., Yildirim, J., and Connor, P. M. (2015). Regional distribution of photovoltaic deployment in the uk and its determinants: a spatial econometric approach. *Energy Economics*, 51:417–429.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling for spatial data*. Chapman and Hall CRC, Boca Raton, FL.

Bayer, P., Ross, S. L., and Topa, G. (2008). Place of work and place of residence: informal hiring networks and labor market outcomes. *Journal of Political Economy*, 116(6):1150–1196.

Ben-Joseph, E. (1995). Residential street standards and neighborhood traffic control: a survey of cities' practices and public officials' attitudes. Working Paper UCB-ITS-WP-95-1. Institute of Transportation Studies, University of California, Berkeley.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–236.

Bollinger, B. and Gillingham, K. (2010). Environmental preferences and peer effects in the diffusion of solar photovoltaic panels. Working paper.

Bollinger, B. and Gillingham, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Marketing science*, 31(6):900–912.

Borck, R. (2007). Consumption and social life in cities: evidence from Germany. *Urban Studies*, 44(11):2105–2121.

Borenstein, S. (2017). Private net benefits of residential solar pv: The role of electricity tariffs, tax incentives, and rebates. *Journal of the Association of Environmental and Resource Economists*, 4(S1):S85–S122.

Brown, B. B. and Cropper, V. L. (2001). New urban and standard suburban subdivisions: evaluating psychological and social goals. *Journal of the American Planning Association*, 67(4):402–419.

Brueckner, J. K. and Largey, A. G. (2008). Social interaction and urban sprawl. *Journal of Urban Economics*, 64(1):18–34.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*. Econometric Society Monographs. Cambridge University Press, New York, NY, 2nd edition.

Cao, X., Handy, S. L., and Mokhtarian, P. L. (2008). Neighborhood design and children's outdoor play: evidence from Northern California. *Children, Youth and Environments*, 18(2):160–179.

Cao, X., Handy, S. L., and Mokhtarian, P. L. (2009). Examining the impacts of residential self-selection on travel behavior: A focus on empirical findings. *Transport Reviews*, 29(3):359–395.

Cao, X. and Mokhtarian, P. L. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B*, 42(3):204–228.

Castro, M., Paleti, R., and Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: application to predicting crash frequency at intersections. *Transportation Research B*, 46(1):253–272.

Charles, K. K. and Kline, P. (2006). Relational costs and the production of social capital: evidence from carpooling. *The Economic Journal*, 116(511):581–604.

Chib, S. and Carlin, B. P. (1999). On mcmc sampling in hierarchical longitudinal models. *Statistics and Computing*, 9(1):17–26.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Chib, S., Greenberg, E., and Winkelmann, R. (1998). Posterior simulation and bayes factors in panel count data models. *Journal of Econometrics*, 86(1):33–54.

Chib, S. and Jeliazkov, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Associations*, 101(474):685–700.

Crago, C. L. and Chernyakhovskiy, I. (2017). Are policy incentives for solar power effective? evidence from residential installations in the northeast. *Journal of Environmental Economics and Management*, 81:132–151.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley and Sons, New York, NY.

CSI (2017). *California Solar Initiative Program Handbook*. California Public Utilities Commission.

Drury, E., Miller, M., Macal, C. M., Graziano, D. J., Heimiller, D., Ozik, J., and IV, T. D. P. (2012). The transformation of southern california's residential photovoltaics market through third-party ownership. *Energy Policy*, 42:681–690.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217.

Elhorst, J. P., Heijnen, P., Samarina, A., and Jacobs, J. P. A. M. (2017). Transitions at different moments in time: A spatial probit approach. *Journal of Applied Econometrics*, 32(2):422–439.

Franzese, R. J., Hays, J. C., and Cook, S. J. (2016). Spatial- and spatiotemporal-autoregressive probit models of interdependent binary outcomes. *Political Science Research and Methods*, 4(1):151–173.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Gillingham, K. and Graziano, M. (2015a). Spatial patterns of solar photovoltaic system adoption: the influence of neighbors and the built environment. *Journal of Economic Geography*, 15(4):815–839.

Gillingham, K. and Graziano, M. (2015b). Spatial patterns of solar photovoltaic system adoption: the influence of neighbors and the built environment. *Journal of Economic Geography*, 15(4):815–839.

Glaeser, E. L. (2000). The future of urban research: nonmarket interactions. *Boorkings-Wharton Papers on Urban Affairs*, 1:101–150.

Glaeser, E. L. and Gottlieb, J. D. (2006). Urban resurgence and the consumer city. *Urban Studies*, 43(8):1275–1299.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, 7th edition.

Hellerstein, J. K., Kutzbach, M. J., and Neumark, D. (2014). Do labor market networks have an important spatial dimension? *Journal of Urban Economics*, 79:39–58.

Hellerstein, J. K., McInerney, M., and Neumark, D. (2011). Neighbors and coworkers: the importance of residential labor market networks. *Journal of Labor Economics*, 29(4):659–695.

IRENA (2018). Renewable capacity statistics. Technical report, International Renewable Energy Agency (IRENA), Abu Dhabi.

Jaffe, A. B. and Stavins, R. N. (1994). The energy paradox and the diffusion of conservation technology. *Resource and Energy Economics*, 16(2):91–122.

Kamruzzaman, M., Wood, L., Hine, J., Currie, G., Giles-Corti, B., and Turrell, G. (2014). Patterns of social capital associated with transit oriented development. *Journal of Transport Geography*, 35:144–155.

Kirkpatrick, A. J. and Bennear, L. S. (2014). Promoting clean energy investment: An empirical analysis of property assessed clean energy. *Journal of Environmental Economics and Management*, 68(2):357–375.

Lambert, D. M., Brown, J. P., and Florax, R. J. (2010). A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. *Regional Science and Urban Economcis*, 40(4):241–252.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Statistics: textbooks and monographs. Taylor and Francis Group, Boca Raton, FL.

Liesenfeld, R., Richard, J.-F., and Vogler, J. (2016a). Likelihood-based inference and prediction in spatio-temporal panel count models for urban crimes. *Journal of Applied Econometrics*.

Liesenfeld, R., Richard, J.-F., and Vogler, J. (2016b). Likelihood evaluation of high-dimensional spatial latent gaussian models with non-gaussian response variables. In Baltagi, B. H., LeSage, J. P., and Pace, R. K., editors, *Spatial Econometrics: Qualitative and Limited Dependent Variables*, volume 37 of *Advances in Econometrics*, pages 35–77. Emerald Group Publishing Limited.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press.

Manski, C. F. (1993). Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Mason, S. and Fredericksse, E. (2011). Fostering neighborhood viscosity: does design matter? *Community Development Journal*, 46(1):7–26.

Mason, S. G. (2010). Can community design build trust? a comparative study of design factors in Boise, Idaho neighborhoods. *Cities*, 27(6):456–465.

Matisoff, D. C. and Johnson, E. P. (2017). The comparative effectiveness of residential solar incentives. *Energy Policy*, 108:44–54.

Mayo, Jr., J. M. (1979). Suburban neighboring and the cul-de-sac street. *Journal of Architectural Research*, 7(1):22–27.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Noll, D., Dawes, C., and Rai, V. (2014). Solar community organizations and active peer effects in the adoption of residential pv. *Energy Policy*, 67:330–343.

Ostrom, E. (1990). *Governing the Commons: the Evolution of Institutions for Collective Action*. The political economy of institutions and decisions. Cambridge University Press.

Pace, R. K. and Barry, R. P. (1997). Quick computation of spatial autoregressive estimators. *Geographical Analysis*, 29(3):232–246.

Parent, O. and LeSage, J. P. (2008). Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. *Journal of Applied Econometrics*, 23:235–256.

Putnam, R. D. (1994). *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, 6(1):65–78.

Putnam, R. D. (2000). *Bowling Alone: the Collapse and Revival of American Community*. Simon and Schuster, New York.

Rai, V., Reeves, D. C., and Margolis, R. (2016). Overcoming barriers and uncertainties in the adoption of residential solar pv. *Renewable Energy*, 89:498–505.

Rai, V. and Robinson, S. A. (2013). Effective information channels for reducing costs of environmentally- friendly technologies: evidence from residential pv markets. *Environmental Research Letters*, 8(1):014044.

Rai, V. and Sigrin, B. (2013). Diffusion of environmentally-friendly energy technologies: buy versus lease differences in residential pv markets. *Environmental Research Letters*, 8(1):014022.

Reeves, D. C. and Rai, V. (2018). Strike while the rebate is hot: Savvy consumers and strategic technology adoption timing. Available at SSRN: https://ssrn.com/abstract=3119133.

Richter, L.-L. (2013). Social effects in the diffusion of solar photovoltaic technology in the uk. *Cambridge Working Paper in Economics*, 1357.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Number 104 in Monographs on statistics and applied probability. Chapman and Hall CRC, Boca Raton, FL.

Sigrin, B., Pless, J., and Drury, E. (2015). Diffusion into new markets: evolving customer segments in the solar photovoltaics market. *Environmental Research Letters*, 10(8):084001.

Simões, P. and Natário, I. (2016). Spatial econometric approaches for count data: an overview and new directions. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 10(1).

Vaishnav, P., Horner, N., and Azevedo, I. L. (2017). Was it worthwhile? where have the benefits of rooftop solar photovoltaic generation exceeded the cost? *Environmental Research Letters*, 12(9):094015.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121:311–324.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, 69(3):309–312.

Wood, L., Giles-Corti, B., and Bulsara, M. (2012). Streets apart: does social capital vary with neighborhood design? *Urban Studies Research*.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1):39–54.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts, 2nd edition.

Appendix A

A.1. Bivariate probit model

Log-likelihood function

The joint probabilities of (C, I), as presented by Greene (2012, p. 747), are

$$P_{11} = Pr(S = 1, D = 1) = \Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho) \tag{A.1}$$

$$P_{10} = Pr(S = 1, D = 0) = \Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho) \tag{A.2}$$

$$P_{01} = Pr(S = 0, D = 1) = \Phi_2(-\boldsymbol{x_2'\beta} - \gamma, \boldsymbol{x_1'\alpha}, -\rho) \tag{A.3}$$

$$P_{00} = Pr(S = 0, D = 0) = \Phi_2(-\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, \rho), \tag{A.4}$$

where $\Phi_2(.)$ denotes the cumulative distribution function (cdf) of the bivariate standard normal distribution. The signs of the function arguments ensure that the four probabilities sum to one. The log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} D_i S_i \ln(P_{11,i}) + D_i(1 - S_i) \ln(P_{10,i}) + (1 - D_i) S_i \ln(P_{01,i}) + (1 - S_i)(1 - D_i) \ln(P_{00,i}). \tag{A.5}$$

Denoting the vector of model parameters by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \rho)$, the log-likelihood function is maximized with respect to $\boldsymbol{\theta}$ to obtain estimates of the parameters, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\rho})$.

Partial Effects

Of primary interest is the partial effect of changing the endogenous neighborhood design variable, D, on the probability of carpooling.There are different kinds of partial effects that can be computed. One of which is the effect on the *marginal* probability of carpooling:

$$E(S \mid \boldsymbol{x_2}, \rho = 0, D = 1) - E(S \mid \boldsymbol{x_2}, \rho = 0, D = 0) = \Phi(\boldsymbol{x_2'\beta} + \gamma) - \Phi(\boldsymbol{x_2'\beta}), \tag{A.6}$$

where $\Phi(.)$ denotes the cdf of the univariate standard normal distribution. Another partial effect is that on the *conditional* probability of carpooling:

$$E(S \mid \boldsymbol{x}, \rho, D = 1) - E(S \mid \boldsymbol{x}, \rho, D = 0) = \frac{\Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho)}{\Phi(\boldsymbol{x_1'\alpha})} - \frac{\Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho)}{1 - \Phi(\boldsymbol{x_1'\alpha})} \tag{A.7}$$

where $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$ and $\Phi_2(.)$ denotes the bivariate standard normal cdf.

Partial effects of changes in the covariates can also be derived. The conditional mean functions of the two endogenous variables (D and S) are

$$E(D \mid \boldsymbol{x_1}) = \Phi(\boldsymbol{\alpha' x_1}) \tag{A.8}$$

$$E(S \mid \boldsymbol{x}) = \Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho) + \Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho) \tag{A.9}$$

For a continuous covariate, x_c, that appears in $\boldsymbol{x_1}$ and/or $\boldsymbol{x_2}$, the partial effect is

$$\begin{aligned}\frac{\partial E(S \mid \boldsymbol{x})}{\partial x_c} = &\left[\phi(\boldsymbol{x_2'\beta} + \gamma)\Phi\left(\frac{\boldsymbol{x_1'\alpha} - \rho(\boldsymbol{x_2'\beta} + \gamma)}{\sqrt{1 - \rho^2}}\right) + \phi(\boldsymbol{x_2'\beta})\Phi\left(\frac{-\boldsymbol{x_1'\alpha} + \rho\boldsymbol{x_2'\beta}}{\sqrt{1 - \rho^2}}\right)\right]\beta_c \\ &+ \left[\phi(\boldsymbol{x_1'\alpha})\Phi\left(\frac{\boldsymbol{x_2'\beta} + \gamma - \rho\boldsymbol{x_1'\alpha}}{\sqrt{1 - \rho^2}}\right) - \phi(-\boldsymbol{x_1'\alpha})\Phi\left(\frac{\boldsymbol{x_2'\beta} - \rho\boldsymbol{x_1'\alpha}}{\sqrt{1 - \rho^2}}\right)\right]\alpha_c \end{aligned} \tag{A.10}$$

where $\phi(.)$ is the pdf of the univariate standard normal distribution, and α_c and β_c are the

coefficients corresponding to the covariate of interest, x_c. The two terms in $(A.10)$ represent the direct and indirect (through D) effects, respectively.

For a binary covariate, x_b, that appears in $\boldsymbol{x_1}$ and/or $\boldsymbol{x_2}$, the partial effect is

$$E(S \mid \boldsymbol{x}, x_b = 1) - E(S \mid \boldsymbol{x}, x_b = 0) = \Big[\Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho) + \Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho)\Big]_{x_b=1} - \Big[\Phi_2(\boldsymbol{x_2'\beta} + \gamma, \boldsymbol{x_1'\alpha}, \rho) + \Phi_2(\boldsymbol{x_2'\beta}, -\boldsymbol{x_1'\alpha}, -\rho)\Big]_{x_b=0} \quad \text{(A.11)}$$

$(A.11)$ is also used for computing the partial effect of a categorical variable by transforming it into a series of binary variables representing choices between a reference category and each of the remaining categories, respectively. Then the partial effect is obtained for the applicable category for each observation.

Finally, the partial effects of the covariates on the choice of neighborhood design (D) can be derived similarly using the conditional mean function of D $(A.8)$.

A.2. Description of covariates

The first group of covariates represents individual characteristics of students. Age controls for the possibility that students widen their network of friends, with whom they may carpool, as they grow up. A quadratic age term is also included in case this network formation accelerates with age. Gender accounts for the possibility that males may be more active outdoors in the neighborhood such that they are exposed to more social interaction. Parents may also be more inclined to allow their children to carpool with non-relatives if they are older or males. These covariates are not included in the neighborhood design equation because this choice is believed to be made by adult parents or heads of households.

Other individual characteristics include race, ethnicity, and citizenship status to control for possible differences in carpooling behavior based on demographics. Finally, travel time to school (based on shortest driving path using speed limits) is included as students living sufficiently close to school may walk instead of being driven to it. It also captures any travel distance considerations that may factor into the choice of neighborhood design.

The second covariate group is related to the household. Larger households may have a preference for more spacious homes in the suburbs where CDS's are located. Households with more children may prefer the CDS for its perceived benefits mentioned previously. On the other hand, having more students in the household, especially if enrolled in different schools, may hinder parents' ability to chauffeur them to school and, hence, increase their need for other modes of travel, including carpooling. The presence of a retired person or homemaker indicates that person is available to accompany the student to school, thus reducing the need to seek carpooling from neighbors. Although this person may at the same time provide an opportunity for other students in the neighborhood to carpool with them, the student under consideration is considered a member of the same household as the retired or homemaker person and would have been coded as an "auto passenger" by the survey administrator.

Household income is expected to be positively correlated with CDS as wealthier households are known to have a preference for and can afford living in the suburbs. However, income can have two offsetting effects on carpooling. First, wealthier households may have less need for carpooling because they can afford to buy cars or have the time to spare on chauffeuring their kids to school. Conversely, from a social interaction standpoint, they may have the time and resources to engage in neighborhood social activities. Along with income, the number of workers in the household is also included to capture the general economic condition of the household and how it relates to the choice of neighborhood design and carpooling.

Tenure represents the number of years the household has lived in their current home. Households with higher tenure are expected to have had more time to form their social networks in the neighborhood than new comers. It is worth mentioning here that home

ownership is not included in the model due to its potentially endogeneity (as BL2008 explain, homeowners may value investments in their neighborhood, in terms of establishment of social bonds with neighbors, more than temporary renters). Its effect, nevertheless, may be captured through tenure.

The number of vehicles in the household is expected to correlate positively with the CDS, where the need for auto travel is higher and private parking space is abundant. The availability of vehicles is also expected to discourage carpooling. The average travel time to work for households with one or more workers is included based on the possibility that it might interfere with the school transportation mode choice.

Additional variables pertaining to the householder (defined by the US Census as the person who owns or rents the property, or self-identified as the head of household in response to the CHTS[2]) that may explain the choice of neighborhood design.

The third covariate group account for the general social environment in the wider area where the individual lives–the county. The racial and ethnic dissimilarity are measures of segregation. A high dissimilarity value indicates a more segregated area where an individual is more exposed to other people of their own race or ethnicity than one in a less segregated area. These variables account for the possibility that individuals may be more likely to socialize with others of their own race or ethnicity. The county murder rate is meant to capture the perception of safety in the county and how it correlates with the neighborhood types. It is also expected to be negatively correlated with carpooling as people may be less trusting of one another.

Finally, the last group of covariates is a set of dummy variables indicating which of the 9 California regions the individual lives in. It is meant to capture unobserved effects in the region such as the physical environment, urban planning and transportation policies, general economic conditions, or social and cultural norms.

[2]Daigler, Vivian (NuStats Research Solutions). "Re: CHTS 2012." Received on 8 Aug. 2017.

A.3. Results from the replication of BL2008

Table A.1: Coefficients from model of census tract density and carpooling to work

Explanatory variables	**Dependent variables**			
	Census tract density[a]		Carpool to work	
	Coefficient	(SE)	APE	(SE)
Census tract density	-	-	−0.0003	(0.0004)
Individual Characteristics				
Age	0.0644	(0.0161)***	0.0000	(0.0001)
Age^2	−0.0008	(0.0002)***	-	-
Male	0.3216	(0.0643)***	0.0021	(0.0015)
Black[b]	1.3684	(0.2581)***	−0.0054	(0.0039)
American Indian/Alaskan Native[b]	0.4324	(0.2510)*	0.0095	(0.0051)*
Asian[b]	0.4994	(0.2437)**	0.0093	(0.0037)**
Mixed/other race[b]	0.3080	(0.1727)*	−0.0009	(0.0027)
Hispanic	0.7784	(0.1587)***	0.0009	(0.0028)
Citizen or in 10+ years	−0.7786	(0.4320)*	−0.0009	(0.0060)
College degree	−0.0606	(0.0980)	−0.0007	(0.0023)
Only worker in the household	0.1790	(0.1376)	−0.0059	(0.0023)***
Travel time to work	−0.0413	(0.0023)***	0.0007	(0.0000)***
Household Characteristics				
Household size	0.4455	(0.0836)***	0.0053	(0.0012)***
Number of children	−0.6210	(0.0982)***	−0.0055	(0.0017)***
Household has students	−0.2969	(0.1643)*	−0.0015	(0.0031)
Retired person/homemaker aged 18-65 in household	−0.4393	(0.1465)	−0.0016	(0.0029)
Income(2)[b] $25,000 - $49,999	−0.6259	(0.2486)**	−0.0006	(0.0034)
Income(3)[b] $50,000 - $99,999	−1.1385	(0.2298)***	0.0005	(0.0033)
Income(4)[b] $100,000 - $149,999	−1.6932	(0.2358)***	0.0058	(0.0038)
Income(5)[b] $150,000 or more	−2.2598	(0.2476)***	0.0025	(0.0041)
Income missing[b]	−1.8579	(0.2724)***	0.0030	(0.0044)
Tenure	0.0541	(0.0128)***	0.0003	(0.0001)***
Tenure^2	−0.0010	(0.0003)***	-	-
Number of vehicles	−1.3378	(0.0787)***	−0.0073	(0.0015)***
Household average travel time to school	−0.0160	(0.0099)	−0.0002	(0.0002)
Age of householder	−0.0580	(0.0288)**	−0.0004	(0.0001)***
Age^2 of householder	0.0001	(0.0003)	-	-
Male householder	0.2594	(0.0946)***	−0.0000	(0.0018)
Householder lives with spouse or partner	−0.3670	(0.1405)***	−0.0002	(0.0025)
Householder has college degree	−0.2122	(0.1190)*	−0.0031	(0.0025)
Householder unemployed[b]	−0.0469	(0.3197)	−0.0057	(0.0045)
Householder retired/homemaker[b]	0.1913	(0.1535)	−0.0029	(0.0034)
Householder other employment status[b]	−0.3203	(0.2902)	−0.0060	(0.0039)
County characteristics				
Racial dissimilarity (white vs. non-white)	6.0565	(0.8850)***	0.0152	(0.0236)
Ethnic dissimilarity (hispanic vs. non-hispanic)	−4.6729	(0.7756)***	−0.0282	(0.0154)*
Murder rate	0.5138	(0.0200)***	0.0004	(0.0004)
Regions[b]				
(2) Central Coast	−1.8628	(0.1749)***	0.0005	(0.0052)
(4) Greater Sacramento	−1.4748	(0.1473)***	−0.0047	(0.0037)
(5) Northern California	0.1841	(0.3562)	−0.0051	(0.0138)
(6) Northern Sacramento Valley	−2.5072	(0.2231)***	−0.0141	(0.0047)***
(7) San Joaquin Valley	−4.1000	(0.1706)***	−0.0053	(0.0041)
(8) Southern Border	0.2952	(0.1662)*	−0.0019	(0.0044)
(9) Southern California	−0.8954	(0.1204)***	0.0034	(0.0030)
Instruments				
MSA population density (in units of 1000)	0.4044	(0.0574)***	-	-
Urban population density (in units of 1000)	1.3866	(0.0298)***	-	-
MSA terrain ruggedness	0.0652	(0.0069)***	-	-
Intercept	2.6443	(0.9100)***	−2.4336	(0.3554)***

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)
$N = 33,140$; number of clusters $= 22,838$.
Goodness of fit: Wald $\chi^2(44)$ statistic (P-value) $= 505.2$ [0.000].
Correlation, $\rho = 0.028$; exogeneity test ($h_o : \rho = 0$) Wald statistic [P-value] $= 0.200$ [0.652].

[a] First-stage: $R^2 = 0.295$; $F(d.f.) = 1041\ (3, 22837)$; Overidentifying restrictions test J-statistic (P-value)= 4.545 [0.103].

[b] Reference categories: Income ($< \$25,000$); race (white); employment status(employed); region (Bay Area)

Table A.2: The effect of census tract density on carpooling to school

Explanatory variables	**Dependent variables**			
	Census tract density[a]		Carpool to school	
	Coefficient	(SE)	APE	(SE)
Census tract population density (1000 people per sqm.)[b]	-	-	−0.0030	(0.0012)**
Individual Characteristics				
Age	0.0976	(0.1526)	0.0030	(0.0009)***
Age^2	−0.0067	(0.0081)	-	-
Male	0.1323	(0.1127)	−0.0026	(0.0036)
Black[b]	1.0256	(0.4433)**	−0.0161	(0.0088)
American Indian/Alaskan Native[b]	0.9023	(0.3376)***	0.0057	(0.0104)
Asian[b]	0.2202	(0.3112)	−0.0146	(0.0064)**
Mixed/other race[b]	0.4340	(0.2141)**	0.0150	(0.0079)*
Hispanic	0.7856	(0.1928)***	−0.0190	(0.0063)***
Citizen or in US 10+ years	−0.3529	0.4222	0.0104	(0.0110)
Travel time to school	−0.0123	(0.0108)	0.0012	(0.0002)***
Household Characteristics				
Household size	0.3236	(0.0809)***	0.0059	(0.0025)**
Number of students	−0.3830	(0.1219)***	−0.0078	(0.0042)*
Income(2)[b] $25,000 - $49,999	0.2112	(0.3144)	−0.0037	(0.0084)
Income(3)[b] $50,000 - $99,999	−0.6002	(0.2923)**	−0.0000	(0.0093)
Income(4)[b] $100,000 - $149,999	−1.0586	(0.3322)***	0.0163	(0.0113)
Income(5)[b] $150,000 or more	−1.7595	(0.3488)***	0.0079	0.0115
Income missing[b]	−0.7201	(0.4501)	0.0058	(0.0120)
Tenure	0.0867	(0.0235)***	0.0008	(0.0004)**
Tenure^2	−0.0013	(0.0005)**	-	-
Number of vehicles	−1.2888	(0.1232)***	−0.0111	(0.0042)***
Household average travel time to work	−0.0279	(0.0043)***	−0.0001	(0.0002)
1-worker household[b]	1.0972	(0.4070)***	0.0054	(0.0094)
2-or-more-worker household[b]	1.6011	(0.4520)***	0.0173	(0.0108)
Age of householder	−0.0613	(0.4626)	−0.0001	(0.0003)
Age^2 of householder	0.0004	(0.0005)	-	-
Male householder	0.0464	(0.1605)	−0.0076	(0.0050)
Householder lives with spouse or partner	0.0784	(0.2341)	0.0000	(0.0072)
Householder has college degree or higher	−0.7663	(0.1787)***	0.0083	(0.0004)**
Householder unemployed[b]	0.2765	(0.3523)	0.0093	(0.0139)
Householder retired/homemaker[b]	0.4575	(0.2539)*	−0.0008	(0.0077)
Householder other employment status[b]	0.1996	(0.3694)	−0.0025	(0.0109)
County characteristics				
Racial dissimilarity (white vs. non-white)	6.6716	(1.1855)***	0.1129	(0.0582)*
EThnic dissimilarity (hispanic vs. non-hispanic)	−0.6340	(1.1453)	−0.0156	(0.0410)
Murder rate	0.3610	(0.0310)***	0.0004	(0.0010)
Regions[b]				
(2) Central Coast	−1.4922	(0.3676)***	−0.0023	(0.0126)
(3) Central Sierra	−1.3603	(0.5366)**	0.0171	(0.0289)
(4) Greater Sacramento	−1.1206	(0.2734)***	0.0075	(0.0114)
(5) Northern California	−1.6877	(0.3810)***	−0.0171	(0.0107)
(6) Northern Sacramento Valley	−1.4923	(0.3394)***	−0.0070	(0.0131)
(7) San Joaquin Valley	−3.6564	(0.2750)***	−0.0115	(0.0081)
(8) Southern Border	0.2781	(0.2985)	0.0027	(0.0105)
(9) Southern California	−0.8232	(0.2071)***	0.0115	(0.0077)
Instruments				
MSA population density (in units of 1000)	0.6802	(0.1112)***	-	-
Urban population density (in units of 1000)	1.3191	(0.0579)***	-	-
Intercept	−0.1810	(1.4548)	−2.0533	(0.7311)***

Statistical significance: $*\,(P < 0.10)$; $**\,(P < 0.05)$; $***\,(P < 0.01)$

$N = 9,581$; number of clusters $= 6,264$.

Goodness of fit: Wald $\chi^2(42)$ statistic [P-value] = 168.4 [0.000].

Correlation, $\rho = 0.186$; exogeneity test ($h_o : \rho = 0$) Wald statistic [P-value] = 4.23 [0.040].

[a] First-stage: $R^2 = 0.354$; $F(d.f.) = 436.0\ (2, 6263)$; Overidentifying restrictions test J-statistic [P-value]= 0.014 [0.905].

[b] Reference categories: Income ($< \$25,000$); race (white); employment status (employed); region (Bay Area)

A.4. Results from the bivariate probit model

Table A.3: Structural coefficient estimates from the restricted ($\rho = 0$) work model

Explanatory variables	**Dependent variables**			
	Neighborhood design (D)		Carpool to school (S)	
	Coef.	(SE)	Coef.	(SE)
CDS (length of 150m)	-	-	−0.0233	(0.0521)
Individual Characteristics				
Age	-	-	0.0167	(0.0089)∗
Age^2	-	-	−0.0002	(0.0001)∗∗
Male	-	-	0.0695	(0.0298)∗∗
Black[a]	−0.0997	(0.0698)	−0.1450	(0.1136)
American Indian or Alaskan Native [a]	−0.0571	(0.0602)	0.1492	9.9684)∗
Asian[a]	0.0601	(0.0416)	0.2153	(0.0584)∗∗∗
Mixed or other race[a]	−0.0317	(0.0400)	−0.0226	(0.0592)
Hispanic	−0.0711	(0.0369)∗	0.0695	(0.0539)
Citizen or in US 10+ years	0.2019	(0.0872)∗∗	−0.0352	(0.1174)
College degree or higher	-	-	−0.0581	(0.0368)
Only worker in the household	−0.0255	0.0291	−0.2401	(0.0482)∗∗∗
Travel time to work	−0.0001	(0.0006)	0.0159	(0.0008)∗∗∗
Household Characteristics				
Household size	−0.0026	(0.0169)	-	-
Number of children	0.0168	(0.0235)	−0.0231	(0.0292)
Household has students	0.1010	(0.0453)∗∗	0.0008	(0.0661)
Retired person or homemaker aged 18-65 in household	-	-	0.0292	(0.0527)
Income(2)[a] $25,000 - $49,999	0.1197	(0.0572)∗∗	−0.0153	(0.0781)
Income(3)[a] $50,000 - $99,999	0.2516	(0.0548)∗∗∗	−0.0110	(0.0725)
Income(4)[a] $100,000 - $149,999	0.3070	(0.0593)∗∗∗	0.0689	(0.0774)
Income(5)[a] $150,000 or more	0.3452	(0.0623)∗∗∗	0.0129	(0.0828)
Income missing[a]	0.2843	(0.0680)∗∗∗	0.0171	(0.0928)
Tenure	0.0199	(0.0034)∗∗∗	0.0123	(0.0062)∗∗
Tenure^2	−0.0005	(0.0001)∗∗∗	−0.0002	(0.0002)
Number of vehicles	0.0610	(0.0152)∗∗∗	−0.1102	(0.0238)∗∗∗
Household average travel time to school	0.0006	(0.0023)	−0.0045	(0.0034)
Age of householder	0.0040	(0.0065)	-	-
(Age of householder)^2	0.0000	(0.0001)	-	-
Male householder	0.0362	(0.0235)	-	-
Householder lives w/ spouse/partner	0.0647	(0.0335)∗	0.0050	(0.0002)
Householder has college degree or higher	0.0542	(0.0262)∗∗	-	-
Householder unemployed[a]	0.0632	(0.0730)	-	-
Householder retired or homemaker[a]	−0.0087	(0.0382)	-	-
Householder other employment status[a]	−0.0396	(0.0638)	-	-
County characteristics				
County racial dissimilarity	1.4973	(0.2708)∗∗∗	0.5290	(0.4000)
County ethnic dissimilarity	−0.8468	(0.2098)∗∗∗	−0.6035	(0.3010)∗∗
County murder rate	0.0022	(0.0041)	0.0014	(0.0058)
Regions[a]				
(2) Central Coast	−0.0343	(0.0613)	0.0554	(0.0886)
(3) Central Sierra	−0.6742	(0.1146)∗∗∗	−0.3150	(0.1471)∗∗
(4) Greater Sacramento	−0.1608	(0.0568)∗∗∗	−0.1194	(0.0940)
(5) Northern California	−0.5725	(0.0956)∗∗∗	−0.2541	(0.1371)∗
(6) Northern Sacramento Valley	−0.0999	(0.0925)	−0.5720	(0.1971)∗∗∗
(7) San Joaquin Valley	0.0495	(0.0483)	−0.0529	(0.0732)
(8) Southern Border	0.0751	(0.0545)	−0.0491	(0.0974)
(9) Southern California	0.0618	(0.0377)	0.0931	(0.0558)∗
Instrument				
% of pre-1940 housing units in county subdivision	−3.1453	(0.2003)∗∗∗	-	-
Intercept	−2.1345	(0.2057)∗∗∗	−2.4313	(0.2629)∗∗∗

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)
$N = 33,140$; number of clusters $= 22,838$.
[a] Reference categories: Income ($< \$25,000$); race (white); employment status (employed); region (Bay Area)

Table A.4: Structural coefficient estimates from the restricted ($\rho = 0$) school model

Explanatory variables	**Dependent variables**			
	Neighborhood design (D)		Carpool to school (S)	
	Coef.	(SE)	Coef.	(SE)
CDS (length of 150m)	-	-	0.2249	(0.0855)***
Individual Characteristics				
Age	-	-	−0.1294	(0.0703)*
Age^2	-	-	0.0089	(0.0037)**
Male	-	-	−0.0437	(0.0529)
Black[a]	−0.2377	(0.1510)	−0.4006	(0.2351)*
American Indian or Alaskan Native [a]	−0.0855	(0.1114)	0.0252	(0.1413)
Asian[a]	−0.0469	(0.0884)	−0.3104	(0.1599)*
Mixed or other race[a]	0.0013	(0.0611)	0.1607	(0.0902)*
Hispanic	−0.0319	(0.0607)	−0.2693	(0.1014)***
Citizen or in US 10+ years	-	-	0.1718	(0.2282)
Travel time to school	−0.0015	(0.0029)	0.0134	(0.0028)***
Household Characteristics				
Household size	−0.0630	(0.0365)*	-	-
Number of children	0.1083	(0.0456)**	-	-
Number of students	-	-	−0.0513	(0.0562)
1-worker household	0.2031	(0.1507)	0.0142	(0.1788)
2-or-more-workers household	0.2852	(0.1603)*	0.0889	(0.1893)
Retired person or homemaker aged 18-65 in household	-	-	−0.1586	(0.0979)
Income(2)[a] $25,000 - $49,999	0.1637	(0.0963)*	−0.0829	(0.1492)
Income(3)[a] $50,000 - $99,999	0.2999	(0.0964)***	0.0181	(0.1525)
Income(4)[a] $100,000 - $149,999	0.4757	(0.1050)***	0.2312	(0.1628)
Income(5)[a] $150,000 or more	0.5621	(0.1126)***	0.1407	(0.1731)
Income missing[a]	0.3915	(0.1265)***	0.0867	(0.1844)
Tenure	0.0034	(0.0033)	0.0148	(0.0107)
Tenure^2	-	-	−0.0004	(0.0003)
Number of vehicles	0.0551	(0.0331)*	−0.0716	(0.0505)
Household average travel time to work	0.0024	(0.0014)*	−0.0001	(0.0024)
Age of householder	−0.0143	(0.0145)	−0.0127	(0.0223)
(Age of householder)^2	0.0002	(0.0002)	0.0001	(0.0002)
Male householder	0.0934	(0.0505)*	−0.1200	(0.0696)*
Householder lives w/ spouse/partner	−0.0028	(0.0771)	0.0695	(0.1092)
Householder has college degree or higher	0.0812	(0.0570)	0.1531	(0.0800)*
Householder unemployed[a]	0.0527	(0.1302)	-	-
Householder retired or homemaker[a]	0.0917	(0.0773)	-	-
Householder other employment status[a]	0.1717	(0.1279)	-	-
County characteristics				
Racial dissimilarity	0.4152	(0.5106)	1.7681	(0.8382)**
Ethnic dissimilarity	−0.5168	(0.4261)	−0.6056	(0.6183)
Murder rate	0.0060	(0.0082)	−0.0144	(0.0105)
Regions[a]				
(2) Central Coast	−0.2138	(0.1309)*	0.1397	(0.1893)
(3) Central Sierra	−0.4257	(0.2604)	−0.0187	(0.2901)
(4) Greater Sacramento	−0.0679	(0.1132)	0.1436	(0.1568)
(5) Northern California	−0.6442	(0.2449)***	−0.2144	(0.2640)
(6) Northern Sacramento Valley	−0.0158	(0.1809)	−0.1087	(0.2474)
(7) San Joaquin Valley	−0.0873	(0.0925)	−0.0119	(0.1341)
(8) Southern Border	0.0831	(0.1076)	0.0929	(0.1597)
(9) Southern California	−0.0770	(0.0744)	0.1925	(0.1102)*
Instrument				
% of pre-1940 housing units in county subdivision	−3.0623	(0.3693)***	-	-
Intercept	−1.4468	(0.4158)***	−1.8742	(0.6814)***

Statistical significance: $*$ ($P < 0.10$); $**$ ($P < 0.05$); $***$ ($P < 0.01$)

[a] Reference categories: Income ($< \$25,000$); race (white); employment status (employed); region (Bay Area)

Appendix B

B.1. Basic Bayesian Updates

The proposed MCMC algorithm relies heavily on standard results from Bayesian analysis of the Normal linear regression model, so it helps to summarize them here first. Consider the model:

$$y = X\beta + \varepsilon$$
$$\varepsilon \sim N_N(0\ ,\ \Sigma)$$

where $\Sigma = \sigma^2 I_N$. Assuming the conjugate prior distributions $\beta \sim N(\beta_0, B_0)$ and $\sigma^2 \sim IG(c/2, d/2)$ results in full conditional posterior distributions of standard form. For β, it is $[\beta|\sigma^2, y] \sim N(\hat{\beta}, \hat{B})$ where $\hat{B}$ is the inverse of the sum of the prior and data precisions (i.e. $\hat{B} = \left[B_0^{-1} + X'\Sigma^{-1}X\right]^{-1}$) and $\hat{\beta}$ is the sum of the precision-weighted average of the prior and data means, re-weighted by the posterior variance (i.e. $\hat{\beta} = \hat{B}\left[B_0^{-1}\beta_0 + X'\Sigma^{-1}y\right]$). For σ^2, the full conditional posterior distribution is $[\sigma^2|\beta, y] \sim IG(\hat{c}/2, \hat{d}/2)$ where the shape parameter is updated by the sample size $\hat{c} = c + N$, and the scale parameter is updated by the sum of squared residuals: $\hat{d} = d + (y - X\beta)'(y - X\beta)$.

B.2. Derivation of the Sampling Distributions

Recall from (2.22) that the augmented joint posterior density is:

$$p(Z, a, \theta|y) \propto \left[\prod_{i}^{N} f_{Pois}(y_i|z_i)\, \phi_T(z_i|\bar{z}_i, \bar{\Omega}_i)\right] \phi_N(a|0, D)\, p(\beta)\, p(\sigma_a^2)\, p(\sigma_\varepsilon^2)\, p(\rho)$$

B.2.1. Sampling Z, β, and a in one block

This block is comprised of two sub-blocks: the first jointly samples Z and β marginally of a, and the second samples a conditional on Z and β.

Sampling (Z, β) marginally of a

The latent variables, Z, are sampled conditionally on one another using the posterior distribution:

$$\left[z_i|\{z_j\}_{j\neq i}, \beta, \sigma_a^2, \sigma_\varepsilon^2, y\right] \sim f_{Pois}(y_i|z_i)\, \phi_T(z_i|\bar{z}_i, \bar{\Omega}_i) \equiv f(z_i|.),$$

which is not of standard form; hence, sampling is done via a MH-step for each vector z_i. Denote the current iterate by z_i^c and let τ_z be a tuning parameter. Then for $i =, 1, 2, ..., N$:

- Draw a candidate from the following random-walk proposal density: $z_i^p = z_i^c + N_T(0, \tau_z I_T)$
- Accept z_i^p with probability $\psi_z = min\{\frac{f(z_i^p|.)}{f(z_i^c|.)}, 1\}$

The result is a $(NT\times 1)$ vector Z of latent variables. Note that the above MH (sub)algorithm is run N times in each cycle of the main MCMC algorithm.

To sample β, note that it only appears in (22) in its prior density and the density of Z. Recall from (20) that the distribution of Z marginalized over a is $N_{NT}(X\beta, \Omega)$, which is essentially a linear sub-model with Z as the continuous dependent variable. Thus, the

posterior distribution of β can be obtained directly from standard results from Bayesian analysis of linear models:

$$\left[\beta | Z, \sigma_a^2, \sigma_\varepsilon^2, \rho\right] \sim N_{\text{K}}(\hat{\beta}, \hat{B}),$$

where

$$\hat{B} = \left[B_0^{-1} + X'\Omega^{-1}X\right]^{-1}$$
$$\hat{\beta} = \hat{B}\left[B_0^{-1}\beta_0 + X'\Omega^{-1}Z\right]$$

Sampling a conditionally (Z, β)

The only term in (22) involving a is its own prior density; thus, using it for sampling a does not take into account updates from the data. Instead, the updating can be obtained by sampling a conditionally on Z and β. Recall the unmarginalized latent-variable representation from (14) where $Z \sim \mathcal{N}\big(X\beta\,,\,[QDQ' + \sigma_\varepsilon^2 I_{NT}]\big)$. Since Z and β have already been sampled (i.e. observed), they can be placed on the left-hand side of (14), which results in the following linear model:

$$(Z - X\beta) = Qa + \varepsilon$$
$$\varepsilon \sim N_{NT}(0\,,\, D + \sigma_\varepsilon^2 I_{NT})$$

where now a are the parameters of interest associated with the covariates matrix, Q. Since this model includes all of the parameters, a is sampled from its full conditional distribution, obtained from standard linear model results:

$$\left[a | Z, \beta, \sigma_a^2, \sigma_\varepsilon^2, \rho, \right] \sim N(\hat{a}, \hat{V}_a)$$

where

$$\hat{V}_a = \left[D^{-1} + \frac{Q'Q}{\sigma_\varepsilon^2}\right]^{-1}$$
$$\hat{a} = \frac{\hat{V}_a Q'(Z - X\beta)}{\sigma_\varepsilon^2}$$

B.2.2. Sampling (σ_a^2, ρ) conditionally on a

σ_a^2 and ρ are sampled jointly in two sub-blocks by simply conditioning on a only, which incorporates all the needed updates from "observed" (i.e. sampled) data thus far. In order to leverage the standard linear model results, write a in its autoregressive representation (11) and re-arrange:

$$A\,a = v$$
$$v \sim N_N(0, \sigma_a^2 I_N)$$
$$A = I - \rho W,$$

which is a simple linear sub-model with $(A\,a)$ being the dependent variable and a zero-mean. The posterior density for this sub-model is then:

$$p(\sigma_a^2, \rho|a) \propto \left[(2\pi\sigma_a^2)^{-N/2}\,|A|\;exp\left(\frac{-a'A'Aa}{2\sigma_a^2}\right)\right] p_{\text{IG}}(\sigma_a^2)\; p_{\text{U}}(\rho)$$

where the term in brackets is the likelihood function, representing the Normal density $\phi_N(Aa|0, \sigma_a^2 I_N)$.

Sampling σ_a^2 conditionally on (ρ, a)

With the inverse-Gamma prior assigned for σ_a^2, its posterior distribution is:

$$[\sigma_a^2|\rho, a] \sim IG\left(\frac{c_1 + N}{2}\;,\; \frac{d_1 + a'A'Aa}{2}\right)$$

Sampling ρ conditionally on (σ_a^2, a)

The posterior distribution of ρ is:

$$[\rho|\sigma_a^2, a] \sim \phi_N(Aa|0, \sigma_a^2 I_N)\, p_{\mathrm{U}}(\rho) \equiv f(\rho|.)$$

which is not of standard form; therefore, ρ is sampled via a random-walk MH step as follows:

- Draw a candidate from the following random-walk proposal density: $\rho^p = \rho^c + N(0, \tau_\rho)$
- Accept ρ^p with probability $\psi_\rho = min\{\frac{f(\rho^p|.)}{f(\rho^c|.)}, 1\}$

where ρ^c and ρ^p are the current and proposed iterates, and τ_ρ is the tuning parameter.

B.2.3. Sampling σ_ε^2

Using the linear model from (1.b) and the inverse-Gamma prior for σ_ε^2 results in the following posterior distribution distribution:

$$\sigma_\varepsilon^2|\beta, Z, a, \sim IG\left(\frac{c_2 + NT}{2}, \frac{d_2 + (Z - X\beta - Qa)'(Z - X\beta - Qa)}{2}\right)$$

Appendix C

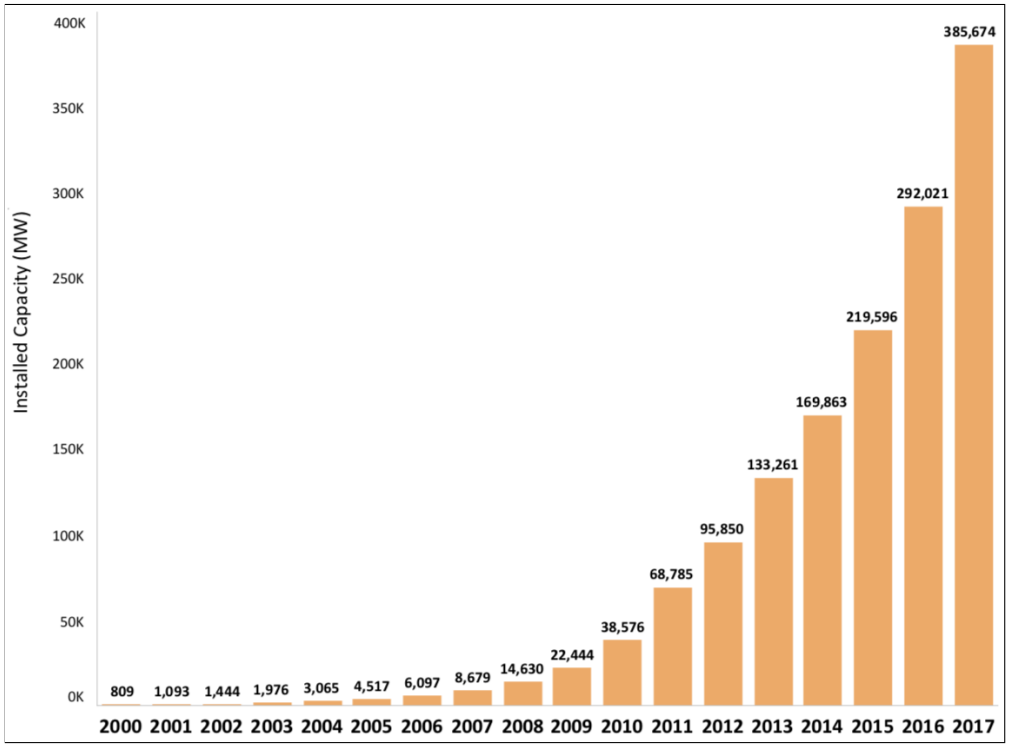

Figure C.1: Global cumulative solar PV installed capacity, 2000-2017. Reprinted from *Solar Energy Data*. Retrieved May 24, 2018, from http://www.irena.org/solar. Copyright 2018 by IRENA. Reprinted with permission.

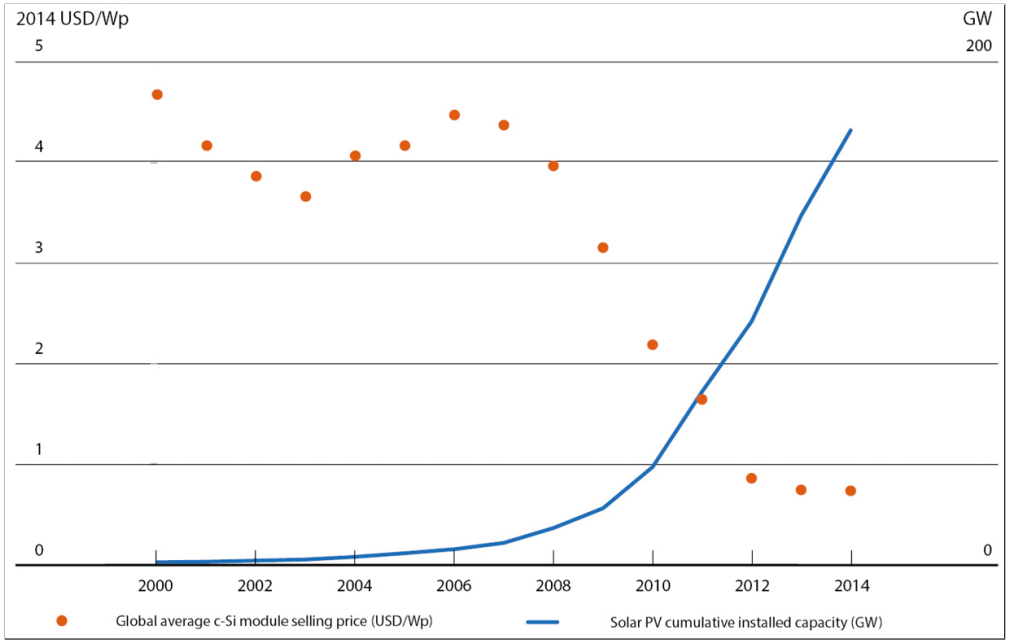

Figure C.2: Cumulative global solar PV deployment and module prices, 2000-2014. Reprinted from *Renewable Power Generation Costs in 2014*. Copyright 2015 by IRENA. Reprinted with permission.

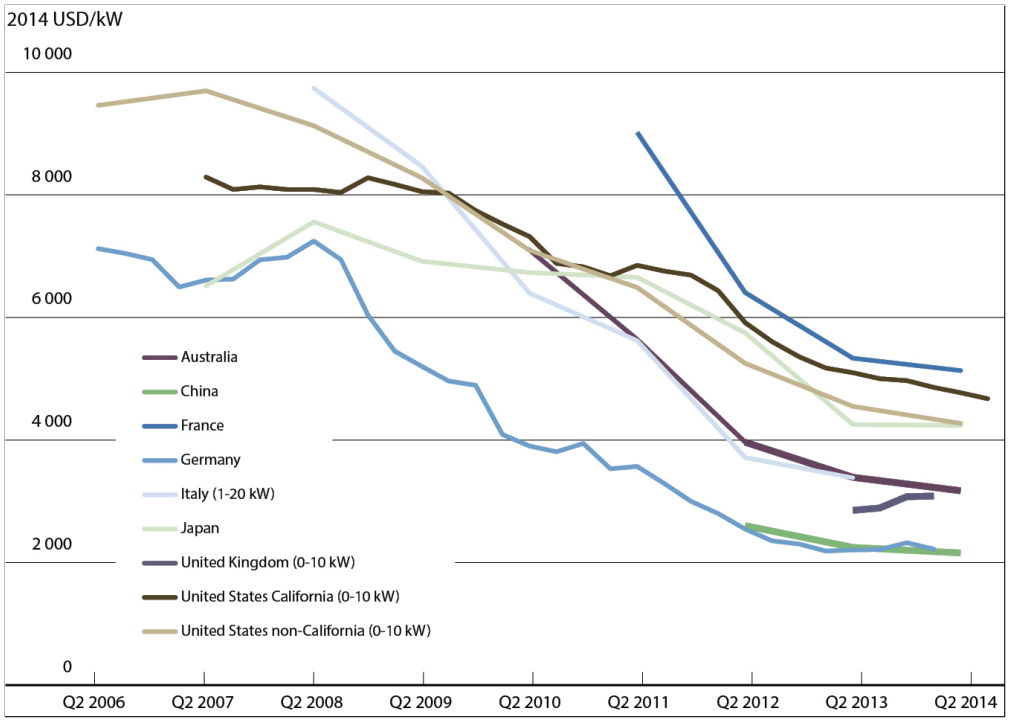

Figure C.3: Average total installed cost of residential solar PV systems by country, 2006-2014. Reprinted from *Renewable Power Generation Costs in 2014*. Copyright 2015 by IRENA. Reprinted with permission.

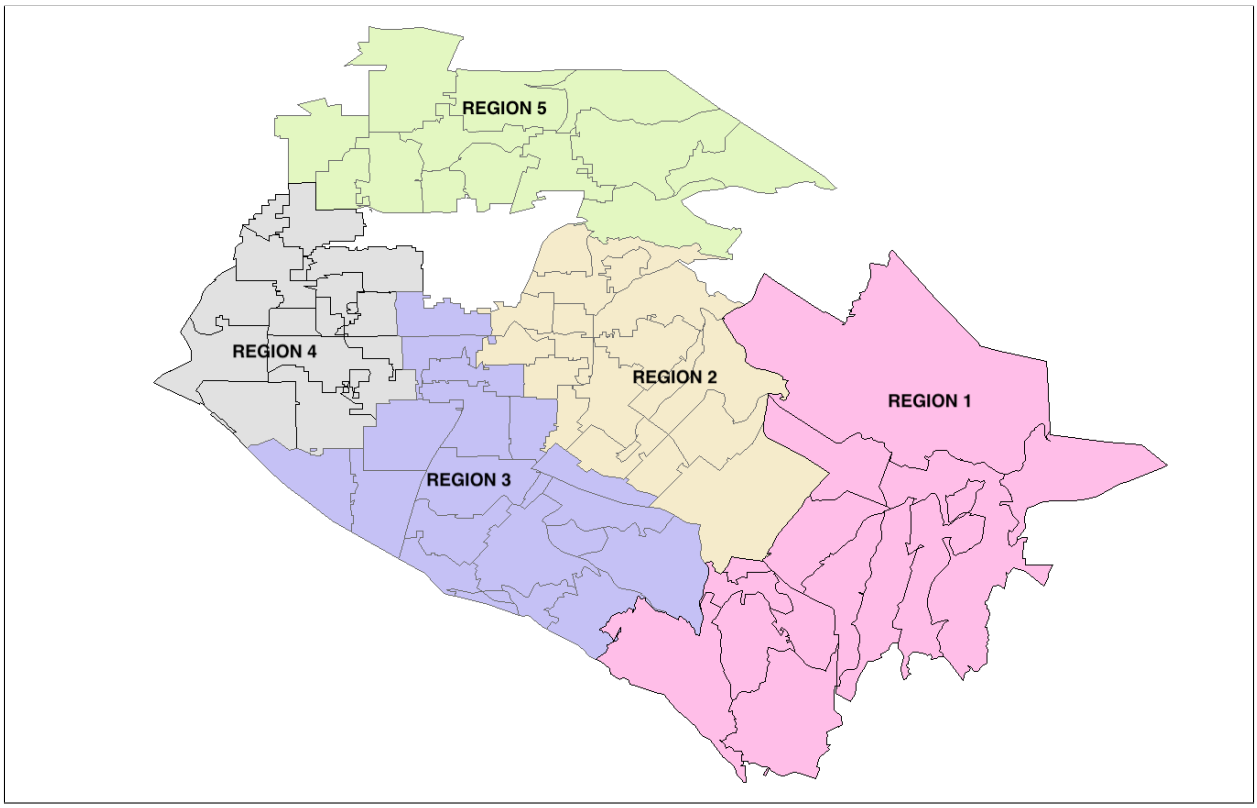

Figure C.4: Orange County