WORKING PAPER SERIES

Estimation of Spillover Effects in Home Mortgage Delinquencies with Sampled Loan Performance Data

Denghui Chen
Office of the Comptroller of the Currency

Hua Kiefer
Federal Deposit Insurance Corporation

Xiaodong Liu
University of Colorado Boulder

March 2021

FDIC CFR WP 2021-03

fdic.gov/cfr

The Center for Financial Research (CFR) Working Paper Series allows CFR staff and their coauthors to circulate preliminary research findings to stimulate discussion and critical comment. Views and opinions expressed in CFR Working Papers reflect those of the authors and do not necessarily reflect those of the FDIC or the United States. Comments and suggestions are welcome and should be directed to the authors. References should cite this research as a “FDIC CFR Working Paper” and should note that findings and conclusions in working papers may be preliminary and subject to revision.
Estimation of Spillover Effects in Home Mortgage Delinquencies with Sampled Loan Performance Data*

Denghui Chen†  Hua Kiefer‡  Xiaodong Liu§

March 16, 2021

Abstract

This paper studies the spillover effect of home mortgage delinquencies using a discrete-choice spatial network model. In our empirical study, a main challenge in estimating this model is that mortgage repayment decisions can only be observed for a sample of all the borrowers in the study region. We show that the nested pseudo-likelihood (NPL) algorithm can be readily modified to accommodate this missing data issue. Monte Carlo simulations indicate that the proposed estimator works well in finite samples and ignoring this issue leads to a downward bias in the estimated spillover effect. We estimate the model using data on single-family residential mortgage delinquencies in Clark County of Nevada in 2010, and find strong evidence of spillover effects. We also conduct some counterfactual experiments to illustrate the policy relevance of the spillover effect.

Keywords: missing data, mortgage defaults, networks, NPL, rational expectation.
JEL: C21, R31

*We thank Ryan Goodstein, Amanda Heitz, Mark Kutzbach, Ajay A. Palvia, Jon Pogach, Alexander Ufier, Ioan Voicu, Chiwon Yom, and the participants of the FDIC CFR seminar for helpful comments. The views expressed in this paper do not necessarily reflect the views of the Federal Deposit Insurance Corporation, the Office of the Comptroller of the Currency, or any federal agency, and do not establish supervisory policy, requirements, or expectations.
†Office of the Comptroller of the Currency, denghui.chen@occ.treas.gov.
‡Federal Deposit Insurance Corporation, hkiefer@fdic.gov.
§Department of Economics, University of Colorado Boulder, xiaodong.liu@colorado.edu.
1 Introduction

In the aftermath of the global financial crisis, the U.S. housing market experienced its worst years since the Great Depression. The number of seriously delinquent loans (i.e., loans that are 90 days past due or worse, aka 90+ DPD) and foreclosures skyrocketed during this period. According to data from the Mortgage Bankers Association, the share of mortgage loans that were 90+ DPD averaged 0.73 percent from 1974 to 2006. By the end of 2009, the rate has surged to 5.09 percent, which means there were 2.26 million stressed mortgage borrowers heading down the road to foreclosure. Considering there were 2 million loans already in the foreclosure process as of the end of 2009 (more than triple the pre-crisis level), adding these stressed borrowers (even only a fraction of them) to the existing pool of foreclosures could lead to a large negative impact on the economy. The negative impact of a foreclosure goes beyond the individual mortgage borrower and lender. Additional costs are incurred by the community (including the nearby homeowners and business owners) and local municipal governments. The magnitude of such costs can be as much as tens of thousand dollars.\(^1\)

In light of the broad reach and steep cost of foreclosures, the large increase of foreclosure activities in the post-crisis period gave rise to a series of federal and proprietary foreclosure prevention programs (e.g., HAMP, HARP, FHA Streamline Refinance, etc.) designed to help financially stressed homeowners save their homes. The foreclosure starts rate has gradually fallen and returned to the pre-crisis level since 2016. However, the recent COVID-19 pandemic has hit the U.S. economy hard and affected many homeowners’ ability to repay their mortgages. In response to the pandemic, Congress passed the CARES Act to provide relief to stressed mortgage borrowers.\(^2\) As a result, the foreclosure inventory rate (i.e., the

---

\(^1\)Cutts & Green (2005) estimate an average cost of $58,792 to the homeowner of the foreclosed property. Moreno (1995) suggests that the cities have to bear a rehab cost ranging from $25,000 to $40,000; or a demolishing cost (for properties beyond repair) between $6,000 and $10,000; and the adjacent homes (two houses on each side and two house directly across the street) suffer a cumulative loss in property value of $10,000. Apgar & Duda (2005) find government costs can exceed $30,000 per property in some cases.

\(^2\)Under the Coronavirus Aid, Relief, and Economic Security (CARES) Act, homeowners with federally or GSE-backed (Fannie Mae or Freddie Mac) or funded mortgages can seek two types of assistance: 1) Foreclosure moratoriums – for federally backed loans, lenders or loan servicers may not foreclose on you until after February 28, 2021; for loans backed by Fannie Mae or Freddie Mac, they cannot foreclose on until
share of mortgages in the foreclosure process) has remained low since the initial onset of the COVID-19. Nevertheless, while we are getting close to the expiration of the CARES Act, the pool size of mortgages in forbearance as of right now is comparable to that of mortgages in serious delinquency back in late 2009,\(^3\) signaling the U.S. could be heading toward a new foreclosure crisis. Therefore, studying home mortgage defaults in the global financial crisis bears great current significance.

In this paper, we adopt the discrete-choice network model in Lee et al. (2014) to study the interdependence of homeowners’ mortgage repayment decisions. In this model, a mortgage borrower’s repayment decision depends not only on neighboring foreclosures in the previous time period (the contagion effect in Towe & Lawley 2013) but also on the expectation of neighbors’ current repayment decisions based on available information (the spillover or multiplier effect in Chomsisengphet et al. 2018). Thus, this model establishes a direct link connecting mortgage repayment decisions of neighboring homeowners to fully capture the spillover effect of mortgage delinquencies.

An underlying assumption in Lee et al. (2014) is that the researcher can observe the outcomes and covariates of all individuals in the network. Although this assumption is quite common for network models, it is not realistic in our empirical study. More specifically, the outcome variable in our model is defined as being 90+ DPD. This information is only available in loan performance data, which is usually collected by the mortgage servicer serving the loans and just covers a portion of all the borrowers in the study region depending on the mortgage servicer’s market share. Ignoring this missing data issue, by treating the sampled borrowers in the loan performance data as the full population, may lead to a biased estimate of the spillover effect. In this paper, we show that, by supplementing the loan performance data with public records on covariates of all the borrowers in the study region, the nested

---

\(^3\) TransUnion reported in its November Industry Snapshot release that 5.85% of the outstanding U.S. mortgages were in hardship (https://www.transunion.com/monthly-industry-snapshot-fs).
pseudo-likelihood (NPL) algorithm (Aguirregabiria & Mira 2007) can be readily modified to accommodate this missing data issue. Our Monte Carlo simulations indicate that the proposed estimator works well in finite samples and ignoring this missing data issue leads to a downward bias in the estimated spillover effect.

In the empirical study, we focus on mortgage borrowers in Clark County of Nevada during a historical foreclosure peak – the year of 2010 – using information from both loan performance data and public records. We find evidence for both a time-lagged contagion effect (Towe & Lawley 2013) and a contemporaneous spillover effect (Chomsisengphet et al. 2018). Consistent with the Monte Carlo simulations, we also find that the spillover effect is underestimated when the missing data problem is left unaddressed. We complement our estimation effort with two counterfactual studies to illustrate the importance of correctly estimated spillover effects in evaluating the effectiveness of a foreclosure prevention program. In the first study, we hypothetically remove properties in foreclosure, one at a time, from the data, and calculate the corresponding reduction in the aggregate delinquency level. In the second study, we introduce a positive utility shock (e.g. a mortgage payment reduction) to all residents in the study region, and plot the percentage reduction in delinquency rates as the shock increases. In both counterfactual studies, we find that the overall reduction in mortgage delinquencies tends to be understated when the contemporaneous spillover effect is ignored or underestimated due to the missing data problem.

Our paper contributes to a large literature on mortgage defaults and corresponding neighborhood effects. One strand of research has shown that mortgage defaults have a significant and highly localized impact on house prices in the neighborhood (Immergluck & Smith 2006, Schuetz et al. 2008, Harding et al. 2009, Campbell et al. 2011, Hartley 2014, Gerardi et al. 2015); while another strand of work has been focusing on the impact of negative equity on default likelihood (Deng et al. 2000, Foote et al. 2008, Bhutta et al. 2010, Elul et al. 2010, Calomiris et al. 2013, Gerardi et al. 2018). Nevertheless, with a couple of exceptions (Towe & Lawley 2013, Chomsisengphet et al. 2018), little work has been done to establish a direct
link among the default decisions of neighboring mortgage borrowers. Towe & Lawley (2013) relate a homeowner’s default decision to the observed default decisions of the neighbors in the previous time period (i.e., neighboring foreclosures). Chomsisengphet et al. (2018) consider a similar framework as the one in this paper, allowing a homeowner’s default decision to depend on the expectation regarding the unobserved default decisions of the neighbors in the current time period. However, Chomsisengphet et al. (2018) have not addressed the missing data problem in the loan performance data.

Our paper also contributes to the literature on partially observed network data. In recent years, substantial progress has been made on the estimation of network effects with partially observed or sampled network data following two research strands. The first strand focuses on partially observed or completely unobserved network links (see, e.g., Liu 2013, Chandrasekhar & Lewis 2016, de Paula et al. 2019, Hardy et al. 2019, Lewbel et al. 2019, Breza et al. 2020, Boucher & Houndetoungan 2020, Griffith 2020), while the second strand focuses on the missing data problem in the outcome or covariates of network nodes (see, e.g., Sojourner 2013, Wang & Lee 2013a,b, Boucher et al. 2014, Liu et al. 2017). In particular, Boucher et al. (2014), Wang & Lee (2013a,b) and Liu et al. (2017) have studied the missing data problem in the outcome variable for linear-in-means models, linear spatial network models, and linear social-interaction models respectively. Our paper complements these studies by addressing this issue in a discrete-choice network model.

The rest of the paper proceeds as follows. Section 2 describes the model, NPL estimation strategy and Monte Carlo simulation experiments. Section 3 presents the data, empirical results and counterfactual studies. Section 4 concludes.


2 Model and NPL Estimation

2.1 Microfoundation

To motivate the econometric specification, we consider the following random utility model. Let \( \mathcal{N} \) denote the set of all borrowers in the study region, with the cardinality denoted by \( n = |\mathcal{N}| \). A borrower’s delinquency decision is indicated by \( y_i \in \{0, 1\} \). As in a standard random utility model, the utility of delinquency \((y_i = 1)\) is normalized to zero, and the utility of making loan payments \((y_i = 0)\) is given by

\[
\epsilon_i - X_i \beta - \lambda \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} y_j,
\]

where \( \epsilon_i \) is an i.i.d. idiosyncratic shock that follows the logistic distribution, \( X_i \) is a row vector of exogenous variables, and \( w_{ij} \) is a normalized spatial weight. More specifically, \( w_{ij} = w_{ij}^*/\sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij}^* \), where \( w_{ij}^* \) is a known constant capturing the geographical proximity between \( i \) and \( j \).\(^4\) Then, \( \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} y_j \) is the distance-weighted delinquency rate in borrower \( i \)'s neighborhood, with its coefficient \( \lambda \) representing the spillover effect of mortgage delinquencies.

As mortgage delinquencies cannot be directly observed by other borrowers, we assume borrowers make delinquency decisions \( y_i \) simultaneously. We further assume that \( X = (X'_1, \cdots, X'_n)' \) and distribution of \( \epsilon_i \) are common knowledge among all borrowers in the area, but the realization of \( \epsilon_i \) is privately observed by borrower \( i \). In the random utility model, borrower \( i \) goes delinquent on loan payments if the expected utility of \( y_i = 0 \), given the information set \( \mathcal{I}_i = \{W, X, \epsilon_i\} \) with \( W = [w_{ij}] \), is less than zero, i.e.

\[
\mathbb{E}(\epsilon_i - X_i \beta - \lambda \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} y_j | \mathcal{I}_i) < 0
\]

\(^4\)In the empirical study, \( w_{ij}^* = 1/d_{ij} \) if \( i \) and \( j \) are within a cutoff distance (say, 0.5 mile), where \( d_{ij} \) denote the geographical distance between \( i \) and \( j \), and \( w_{ij}^* = 0 \) otherwise.
or, equivalently,
\[ X_i\beta + \lambda \sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij}E(y_j|I_i) > \epsilon_i. \]

As \( \epsilon_i \) follows the logistic distribution, borrower \( i \)'s probability of delinquency is
\[ p_i \equiv \Pr(y_i = 1) = \frac{\exp(X_i\beta + \lambda \sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij}E(y_j|I_i))}{1 + \exp(X_i\beta + \lambda \sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij}E(y_j|I_i))}. \]

In the rational expectation equilibrium (Brock & Durlauf 2001a, b), borrower \( i \)'s expectation on borrower \( j \)'s delinquency decision, i.e. \( E(y_j|I_i) \), should be equal to the mathematical probability for borrower \( j \) to be delinquent, i.e. \( p_j \). Therefore, in the equilibrium,
\[ p_i = \Lambda_i(\theta, p), \tag{2} \]
where
\[ \Lambda_i(\theta, p) = \frac{\exp(X_i\beta + \lambda \sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij}p_j)}{1 + \exp(X_i\beta + \lambda \sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij}p_j)} \tag{3} \]
with \( \theta = (\lambda, \beta')' \) and \( p = (p_1, \cdots, p_n)' \). Lee et al. (2014) provide a sufficient condition for the existence of a unique Bayesian Nash equilibrium characterized by Equation (2), which is simply \( |\lambda| < 4 \) in our case. When \( |\lambda| < 4 \), Equation (2) is a contraction mapping and can be solved by recursive iterations.

### 2.2 NPL estimation with sampled loan performance data

The main difficulty in estimating Equation (2) is that \( p = (p_1, \cdots, p_n)' \) is not observable. Lee et al. (2014) suggest to use the nested fixed point (NFXP) algorithm (Rust 1987), with an internal subroutine that solves the fixed point problem given by Equation (2) for \( p \), to implement the maximum likelihood (ML) estimation. To bypass the computational burden of the NFXP algorithm caused by repeatedly solving the fixed point problem at each candidate parameter value in the search for the maximum of the log-likelihood function, Chomsisengphet et al. (2018) adopt the NPL algorithm (Aguirregabiria & Mira 2007) that
is computationally more efficient.

Both Lee et al. (2014) and Chomsisengphet et al. (2018) assume that the researcher can observe the outcomes and exogenous characteristics of all individuals in the network. In our empirical study, the exogenous variables $X_i$ include the home ownership, house square footage, number of bedrooms, property value, loan-to-value (LTV) ratio, and number of foreclosures initiated in the previous period in borrower $i$’s neighborhood. All this information is public and available for all borrowers in disclosure states. On the other hand, the outcome variable $y_i$ is defined as being 90+ DPD, and this information is only available in loan performance data. In the empirical study, we use the loan performance data assembled by a government agency that regulates several national mortgage servicers. Similarly to other popular residential mortgage databases that are commercially available (e.g., CoreLogic or BlackKnight), the coverage of this data depends on the mortgage servicer’s market share. For the specific study region of our empirical analysis, Clark County of Nevada, this data covers about 26% of the single-family residential mortgages. In other words, among all the mortgage repayment decisions in the population, about 26% of them are observed and recorded in our data. Ignoring this missing data issue, by treating the borrowers covered in the loan performance data as the full population, may lead to biased estimates. Although this data issue is inevitable in most cases, it is not unsolvable. In the following, we show that the NPL algorithm can be readily modified to overcome this problem.

Suppose we can observe the exogenous variables $X_i$ for all $i \in \mathcal{N}$, and the delinquency decision $y_i$ for $i \in \mathcal{N}^*$, where $\mathcal{N}^*$ is a random sample of $\mathcal{N}$ with $n^* = |\mathcal{N}^*|$. The NPL algorithm starts from an initial value $p^{(0)} = (p_1^{(0)}, \ldots, p_n^{(0)})'$ and takes the following iterative steps:

---

5 Both property value and LTV ratio are recorded on the loan origination date in the publicly available transaction data. The initiation of foreclosure is indicated by the notice of default (NOD) or the notice of trustee sale (NOTS) filed in the county office, which is also publicly available.

6 In non-disclosure states, including Alaska, Idaho, Kansas, Louisiana, Mississippi, Missouri (some counties), Montana, New Mexico, North Dakota, Texas, Utah and Wyoming, transaction sale prices are not available to the public. However, since our study area is in a disclosure state, all the aforementioned information are publicly observable.
Step 1 Given $p^{(t-1)} = (p_1^{(t-1)}, \ldots, p_n^{(t-1)})'$, obtain $\hat{\theta}^{(t)} = \arg \max L(\theta; p^{(t-1)})$, where

$$\ln L(\theta; p^{(t-1)}) = \sum_{i \in N^*} \{ y_i \ln \Lambda_i(\theta, p^{(t-1)}) + (1 - y_i) \ln [1 - \Lambda_i(\theta, p^{(t-1)})] \},$$

with $\Lambda_i(\theta, p)$ defined in Equation (3).

Step 2 Given $\hat{\theta}^{(t)}$, update $p^{(t)} = (p_1^{(t)}, \ldots, p_n^{(t)})'$ according to

$$p_i^{(t)} = \Lambda_i(\hat{\theta}^{(t)}, p^{(t-1)}).$$

Repeat Steps 1 and 2 until the process converges.

As we can observe $y_i$ only for $i \in N^*$, a random sample of all borrowers in the full population $\mathcal{N}$, $\theta$ is estimated based on the joint distribution of $y_i$ for $i \in N^*$ in Step 1 of the NPL algorithm. Contrastingly, to update the delinquency probability $p_i^{(t)}$ in Step 2 of the NPL algorithm, we use $X_i$ for all $i \in \mathcal{N}$. Hence, to consistently estimate Equation (2) using the NPL algorithm, one needs to supplement the loan performance data with the information on all borrowers from public records. If, instead, one falsely assumes that $\mathcal{N}^*$ contains all borrowers in the study region and only uses the information in the loan performance data to estimate Equation (2), then the equilibrium delinquency probability $p_i$ would be miscalculated, which introduces a measurement error to the weighted average delinquency probability of the neighbors $\sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij}p_j$ in Equation (3).

Kasahara & Shimotsu (2012) show that a key determinant of the convergence of the NPL algorithm is the contraction property of Equation (2), which is ensured by the assumption of $|\lambda| < 4$. When the NPL algorithm converges, the NPL estimator $\hat{\theta}$ satisfies $\hat{\theta} = \arg \max L(\theta; \tilde{p})$, where $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n)'$ are the solution of the system of equations

$$\tilde{p}_i = \Lambda_i(\hat{\theta}, \tilde{p}),$$
for $i = 1, \ldots, n$. Under some standard regularity conditions, it follows by a similar argument as in Aguirregabiria & Mira (2007) and Lin & Xu (2017) that the NPL estimator is consistent and asymptotically normal.

The estimation of the asymptotic variance of the NPL estimator $\hat{\theta}$ also needs to take this missing data issue into consideration. Let $\hat{\Theta}$ be an $n \times n$ diagonal matrix with the $i$th diagonal element being $\Lambda_i(\hat{\theta}, \hat{p})[1 - \Lambda_i(\hat{\theta}, \hat{p})]$. Let $J$ be a $n^* \times n$ selector matrix such that $Jy$ collects elements in $y = (y_1, \ldots, y_n)'$ corresponding to $i \in \mathcal{N}^*$. The asymptotic variance of $\hat{\theta}$ can be estimated by

$$(\hat{\Sigma}_1 + \hat{\lambda}\hat{\Sigma}_2)^{-1}\hat{\Sigma}_1(\hat{\Sigma}_1 + \hat{\lambda}\hat{\Sigma}_2)^{-1}$$

where

$$\hat{\Sigma}_1 = [X, W\hat{p}]'JJ\hat{\Theta}J'[X, W\hat{p}]$$
$$\hat{\Sigma}_2 = [X, W\hat{p}]'JJ\hat{\Theta}J'J[1 - \hat{\lambda}\hat{\Theta}]^{-1}\hat{\Theta}[X, W\hat{p}].$$

2.3 Monte Carlo Simulations

To investigate the finite sample performance of the proposed estimation procedure, we conduct a simulation study. In the data generating process, we consider two spatial layouts based on Rook contiguity and Queen contiguity. More specifically, we allocate $n = 2500$ spatial units into a lattice of $50 \times 50$ squares. Under Rook contiguity, $w^*_{ij} = 1$ if the squares containing $i$ and $j$ share a common side and $w^*_{ij} = 0$ otherwise. Under Queen contiguity, $w^*_{ij} = 1$ if the squares containing $i$ and $j$ share a common side or vertex and $w^*_{ij} = 0$ otherwise. The normalized spatial weight is given by $w_{ij} = w^*_{ij}/\sum_{j \in \mathcal{N}\setminus\{i\}} w^*_{ij}$. We assume, in Equation (3), that $X_i$ only includes a single exogenous variable that is i.i.d. standard normal, and $\lambda = \beta = 1$. We experiment with different sampling rates $n^*/n \in \{1, 0.75, 0.5, 0.25\}$, and assume that $X_i$ is observable for all $i \in \mathcal{N}$ while $y_i$ is observable only for $i \in \mathcal{N}^*$.

We consider two NPL estimators in the simulation study. The NPL-1 estimator falsely
Table 1: Monte Carlo Simulation Results

<table>
<thead>
<tr>
<th></th>
<th>NPL-1</th>
<th></th>
<th>NPL-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda = 1$</td>
<td>$\beta = 1$</td>
<td>$\lambda = 1$</td>
</tr>
<tr>
<td>Rook contiguity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n^*/n = 1.00$</td>
<td>0.998(0.060)</td>
<td>1.001(0.052)</td>
<td>0.998(0.060)</td>
</tr>
<tr>
<td>$n^*/n = 0.75$</td>
<td>0.989(0.074)</td>
<td>1.001(0.060)</td>
<td>1.000(0.074)</td>
</tr>
<tr>
<td>$n^*/n = 0.50$</td>
<td>0.971(0.088)</td>
<td>0.986(0.073)</td>
<td>1.000(0.086)</td>
</tr>
<tr>
<td>$n^*/n = 0.25$</td>
<td>0.950(0.151)</td>
<td>0.976(0.105)</td>
<td>0.996(0.124)</td>
</tr>
<tr>
<td>Queen contiguity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n^*/n = 1.00$</td>
<td>0.998(0.061)</td>
<td>1.001(0.052)</td>
<td>0.998(0.061)</td>
</tr>
<tr>
<td>$n^*/n = 0.75$</td>
<td>0.996(0.073)</td>
<td>1.004(0.059)</td>
<td>1.000(0.073)</td>
</tr>
<tr>
<td>$n^*/n = 0.50$</td>
<td>0.984(0.087)</td>
<td>0.994(0.073)</td>
<td>0.999(0.086)</td>
</tr>
<tr>
<td>$n^*/n = 0.25$</td>
<td>0.957(0.135)</td>
<td>0.988(0.106)</td>
<td>0.997(0.126)</td>
</tr>
</tbody>
</table>

Mean(SD)

It treats the sample $N^*$ as the full population. As a result, the equilibrium delinquency probabilities $p = (p_1, \cdots, p_n)'$ are miscalculated, and so is the average delinquency probability of the neighbors $\sum_{j \in N \setminus \{i\}} w_{ij}p_j$ in Equation (3).7 The NPL-2 estimator follows the estimation procedure described in Section 2.2 and estimate the model with the correct equilibrium delinquency probabilities. We conduct 1000 simulation repetitions. The mean and standard deviation (SD) of the empirical distribution of the NPL estimates are reported in Table 1. With the miscalculated equilibrium delinquency probabilities, the NPL-1 estimate of $\lambda$ is downward biased while the estimate of $\beta$ is essentially unbiased. The bias increases as the sampling rate decreases. The bias is slightly larger when the underlying spatial network is more sparse (under Rook contiguity). On the other hand, with the correct equilibrium delinquency probabilities, the NPL-2 estimates of $\lambda$ and $\beta$ are unbiased, even when the sampling rate is low ($n^*/n = 0.25$).

7It is worth pointing out that when $N^*$ is treated as the full population, the normalized spatial weight is given by $w_{ij} = w^*_{ij}/\sum_{j \in N^* \setminus \{i\}} w^*_{ij}$.
<table>
<thead>
<tr>
<th><strong>Dependent Variable</strong></th>
<th>Definition</th>
<th>RRP data Mean</th>
<th>RRP data SD</th>
<th>MM data Mean</th>
<th>MM data SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>delinquency</td>
<td>1 if 90+ DPD in 2010, and 0 otherwise.</td>
<td></td>
<td></td>
<td>0.19</td>
<td>0.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Explanatory Variables</strong></th>
<th>Definition</th>
<th>RRP data Mean</th>
<th>RRP data SD</th>
<th>MM data Mean</th>
<th>MM data SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>neighbor foreclosures</td>
<td># of foreclosures (NOD and NOTS) initiated in 2009 within the 0.1 mile neighborhood.</td>
<td>16.69</td>
<td>13.68</td>
<td>14.88</td>
<td>12.04</td>
</tr>
<tr>
<td>owner</td>
<td>1 if the property is occupied by the owner.</td>
<td>0.73</td>
<td>0.44</td>
<td>0.78</td>
<td>0.41</td>
</tr>
<tr>
<td>square footage</td>
<td>The property size in thousand square feet.</td>
<td>2.05</td>
<td>0.78</td>
<td>2.04</td>
<td>0.75</td>
</tr>
<tr>
<td>bedrooms</td>
<td># of bedrooms of the property.</td>
<td>3.39</td>
<td>0.83</td>
<td>3.40</td>
<td>0.82</td>
</tr>
<tr>
<td>log property value</td>
<td>The logarithm of the property’s value at the loan origination date.</td>
<td>12.35</td>
<td>0.52</td>
<td>12.27</td>
<td>0.51</td>
</tr>
<tr>
<td>LTV_60to80</td>
<td>1 if the LTV ratio at the loan origination date is between 60% and 80%.</td>
<td>0.28</td>
<td>0.45</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>LTV_80to100</td>
<td>1 if the LTV ratio at the loan origination date is between 80% and 100%.</td>
<td>0.56</td>
<td>0.50</td>
<td>0.52</td>
<td>0.50</td>
</tr>
<tr>
<td>LTV_gt100</td>
<td>1 if the LTV ratio at the loan origination date is greater than 100%.</td>
<td>0.07</td>
<td>0.25</td>
<td>0.07</td>
<td>0.26</td>
</tr>
</tbody>
</table>

| # of observations         | 221,947    | 58,526        | 0.00        | 0.00         | 0.00       |
3 Empirical Analysis

3.1 Data

Our main data sources are the Mortgage Metrics (MM) database and the Renwood Realty Property (RRP) database. The MM data, assembled by the Office of the Comptroller of the Currency (OCC) since January 2008, consists of loan-level origination and monthly performance information of residential first-lien mortgages serviced by seven national banks and a federal savings association regulated by the OCC. The RRP data covers over 151 million properties and 3,143 counties which translates into 99% of the U.S. population coverage.\(^8\)

We focus on the single-family residential mortgage repayment information in the MM data for Clark County of Nevada in 2010. The RRP data provide a wholistic coverage on the covariates of almost all single-family mortgage borrowers in that region, including those not in the MM data. Using the notation in Section 2.2, we consider the set of borrowers in the RRP data as \(N\) and that in the MM data as \(N^*\). In our study region, the RRP transaction data contains 221,947 loan records distributed across 155 census tracts,\(^9\) whereas the MM sample only has 58,526 records.

The MM data is in a panel structure with monthly updated information for loan performance. The outcome variable of the empirical model – mortgage delinquency (90+ DPD) – is extracted from this data. It is worth pointing out that a mortgage delinquency is different from a foreclosure. The former is a decision made by a homeowner to not make a mortgage payment, while the latter is a legal process in which a lender attempts to recover the balance.

\(^8\)The RRP database consists of three types of data: (1) the transaction data, which provides a history of sales and financing activities on residential housing units, (2) the property tax assessment data collected from county (township) tax assessor’s office, and (3) the pre-foreclosure data (e.g., public records of NOD and NOTS). We use mortgage transaction data (excluding cash transactions) in RRP to construct the pool of active mortgages in the study region. Although we do not know if a mortgage is paid off at the time of our analysis, we feel comfortable that our RRP mortgage data provides a reasonable proxy of the true “active” mortgage population given the fact that the average loan age of the mortgages in our study region is 5.2 years as of the end of 2009. We use the RRP tax assessment data for a complete set of housing characteristic measures. The pre-foreclosure data of RRP provides us information of the existing foreclosure filings, through which we can identify the contagion effect.

\(^9\)We focus on census tracts where most single family homes are located by dropping census tracts with less than 1000 single-family loan records in the RRP data.
of a loan from a borrower who has stopped making payments to the lender by forcing the
sale of the asset used as the collateral for the loan. Once a loan reaches a serious delin-
quency state, such as 90+ DPD, it is usually up to the state level laws and policies (e.g., the
foreclosure law) as well as financial institute level programs (e.g., proprietary modification
programs for loss mitigation) to determine how the foreclosure process proceeds. Because we
are interested in a borrower’s decision instead of the legal aspect of its consequence, we define
the outcome variable as being 90+ DPD in 2010. On the other hand, it is well documented
that mortgage delinquency decisions could be affected by neighboring foreclosures in the
previous time period (Towe & Lawley 2013). Hence, we include the number of foreclosures
initiated in 2009 in a borrower’s neighborhood as a covariate in the empirical model. The
initiation of foreclosure is indicated by the notice of default (NOD) or the notice of trustee
sale (NOTS) filed in the county office. This information is publicly available in a disclosure
state (e.g., Nevada) and contained in the RRP data. Other covariates in the empirical model
includes the home ownership, house square footage, number of bedrooms, property value (on
the loan origination date), and LTV ratio (on the loan origination date). All this information
is also publicly available and contained in the RRP data. We match the loans in the MM
data with those in the RRP data based on encrypted property IDs.\textsuperscript{10}

Table 2 lists the definitions of the dependent variable and explanatory variables as well
as their summary statistics for both the RRP and MM datasets. Overall, the summary
statistics of the explanatory variables are comparable between the two datasets. In both
datasets, the average number of neighbor foreclosures is 15~17. The majority of mortgage
borrowers claimed to be the owners of their properties. The average size of the property
is 2000 square feet, and the typical number of bedrooms is between 3 and 4. The average
property value is about $220K. The number of borrowers with an initial LTV greater than
80\% is slightly more than the number of borrowers with an initial LTV less than 80\%.

\textsuperscript{10}The encrypted property IDs were generated based on the actual address of each property. After the
encrypted property IDs were generated, address information has been removed from both the RRP and MM
data. Thus we, as the end data user, have no access to personally identifiable information.
In the empirical analysis, we treat a census tract as a spatial network. Thus, the scope of spatial interactions is restricted to the census tract level. It is natural to think that the degree of proximity between houses affects the extent of interacted delinquency decisions. We therefore use a normalized spatial weight $w_{ij} = w_{ij}^* / \sum_{j \in N \setminus \{i\}} w_{ij}^*$, where $w_{ij}^*$ depends on the geographical proximity between houses $i$ and $j$, to capture the spatial relationship between houses. The literature suggests that the spillover effect of distressed properties is very local, and this effect decays rapidly with distance (e.g., Campbell et al. 2011, Gerardi et al. 2015, Cohen et al. 2016). We therefore adopt the conventional inverse-distance-based spatial weights and assign zero weights to houses located farther than a cutoff-distance apart. More specifically, $w_{ij}^* = 1/d_{ij}$ if $i$ and $j$ are within a cutoff distance, where $d_{ij}$ denote the geographical distance between $i$ and $j$, and $w_{ij}^* = 0$ otherwise. In the empirical study, we experiment with different cutoff distances ranging from 0.5 miles to 0.1 miles and find the estimation results are robust. Figures 1-3 give a visualization of the average number of neighbors of each house with different cutoff distances for the census tracts used in the empirical analysis. Figure 4 plots the distribution of the delinquency rate in each house’s neighborhood with different cutoff distances using the MM data. We can see that the distribution of the delinquency rate is quite stable with different cutoff distances.

3.2 Estimation Results

The estimation results are reported in Table 3. The first column reports the standard logit estimates without accounting for the delinquency spillover effect. The second and third columns report the NPL estimates of Equation (2) with the delinquency spillover effect. As discussed in Section 2.2, to update the equilibrium delinquency probability $p_i$ in the NPL algorithm, NPL-1 falsely treat the borrowers in the MM data as the whole population and only uses the information on those borrowers and their properties to calculate $p_i$, while NPL-2 uses the information on all borrowers and their properties in the RRP data to calculate $p_i$. In order to control for unobserved regional heterogeneity (e.g., regional demographic and
Figure 1: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.5 miles
Figure 2: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.25 miles
Figure 3: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.1 miles
Figure 4: Distribution of Delinquency Rates with Different Cutoff Distances
economic conditions), sorting of similar borrowers into a neighborhood, and regional random shocks (e.g., regional layoffs) that could be confounded with the delinquency spillover effect, we include block group fixed effects in the estimation.\footnote{A block group is a subdivision of a census tract or block numbering area. It is the smallest geographic entity for which the decennial census tabulates and publishes sample data.}

For the logit model, all the coefficient estimates are statistically significant at the 5% level with the expected signs (except that of \textit{bedroom} is statistically insignificant). In particular, a borrower’s delinquency risk increases with more neighboring foreclosures in the previous time period, giving evidence to the contagion effect (Towe & Lawley 2013). The delinquency risk also increases with a higher property value and LTV ratio. After controlling for the other covariates (including the property value), larger houses have lower delinquency risks. The positive sign of the coefficient estimate of \textit{owner} is not surprising since occupancy fraud is found to be common in various mortgage markets, including government-sponsored-enterprise-guaranteed, private-securitized, and portfolio-held mortgage markets (Haughwout et al. 2011, Elul & Tilson 2015, Piskorski et al. 2015, Griffin & Maturana 2016). Both Haughwout et al. (2011) and Griffin & Maturana (2016) suggest loans with fraud occupancy status perform much worse than otherwise comparable loans.

In both the NPL-1 and NPL-2 models, we find a positive and significant delinquency spillover effect, while the coefficient estimates for other covariates remain largely the same. The NPL-2 estimate of the delinquency spillover effect is more than twice the NPL-1 estimate. This is consistent with our findings in the Monte Carlo simulations that ignoring the missing data issue in the MM data leads to a substantial downward bias of the estimated spillover effect.

In our specification of the spatial weight $w_{ij}$, we assume $w_{ij}$ is zero if the geographical distance between homeowners $i$ and $j$ is greater than a cutoff distance. For the estimation results reported in Table 3, the cutoff distance is set to 0.5 miles. As a robustness check, Table 4 reports the NPL estimates with a cutoff distance of 0.1, 0.25, and 0.5 miles respectively. The results from this sensitivity analysis are reasonable and consistent with our main findings.
For most covariates, the estimated coefficients are very similar with different cutoff distances. For the delinquency spillover effect, as the cutoff distance decreases,

the NPL-1 estimate decreases significantly (e.g., 1.12 for 0.5 miles and 0.38 for 0.1 miles, with a drop of 66%) whereas the NPL-2 estimate is considerably stable across different cutoffs (2.32 for 0.5 miles and 2.12 for 0.1 mile, with a drop of less than 10%). The high sensitivity of the NPL-1 estimate to the cutoff distance makes intuitive sense. As the spatial network becomes more sparse with a shorter cutoff distance, the bias of the NPL-1 estimate due to the missing data issue becomes more severe since the neighborhood delinquencies in the MM sample become less representative of the population level. This is consistent with our findings in the Monte Carlo simulations that the bias of the estimated spillover effect is larger with a more sparse network (with Rook contiguity).

Table 3: Estimation Results

<table>
<thead>
<tr>
<th></th>
<th>Logit</th>
<th>NPL-1</th>
<th>NPL-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>delinquency spillover</td>
<td></td>
<td>1.1210*</td>
<td>2.3226***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.6459)</td>
<td>(0.4629)</td>
</tr>
<tr>
<td>foreclosure contagion</td>
<td>0.0063***</td>
<td>0.0058***</td>
<td>0.0048***</td>
</tr>
<tr>
<td></td>
<td>(0.0013)</td>
<td>(0.0012)</td>
<td>(0.0011)</td>
</tr>
<tr>
<td>owner</td>
<td>0.0586**</td>
<td>0.0585**</td>
<td>0.0593**</td>
</tr>
<tr>
<td></td>
<td>(0.0274)</td>
<td>(0.0270)</td>
<td>(0.0270)</td>
</tr>
<tr>
<td>square footage</td>
<td>-0.3104***</td>
<td>-0.3084***</td>
<td>-0.3061***</td>
</tr>
<tr>
<td></td>
<td>(0.0260)</td>
<td>(0.0266)</td>
<td>(0.0257)</td>
</tr>
<tr>
<td>bedrooms</td>
<td></td>
<td>0.0268</td>
<td>0.0261</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0197)</td>
<td>(0.0194)</td>
</tr>
<tr>
<td>log property value</td>
<td>0.8200***</td>
<td>0.8149***</td>
<td>0.8062***</td>
</tr>
<tr>
<td></td>
<td>(0.0289)</td>
<td>(0.0325)</td>
<td>(0.0325)</td>
</tr>
<tr>
<td>LTV_60to80</td>
<td>0.4499***</td>
<td>0.4499***</td>
<td>0.4487***</td>
</tr>
<tr>
<td></td>
<td>(0.0496)</td>
<td>(0.0495)</td>
<td>(0.0494)</td>
</tr>
<tr>
<td>LTV_80to100</td>
<td>0.8105***</td>
<td>0.8103***</td>
<td>0.8080***</td>
</tr>
<tr>
<td></td>
<td>(0.0479)</td>
<td>(0.0479)</td>
<td>(0.0479)</td>
</tr>
<tr>
<td>LTV_gt100</td>
<td>0.6670***</td>
<td>0.6659***</td>
<td>0.6630***</td>
</tr>
<tr>
<td></td>
<td>(0.0613)</td>
<td>(0.0617)</td>
<td>(0.0617)</td>
</tr>
<tr>
<td>block group dummies</td>
<td>included</td>
<td>included</td>
<td>included</td>
</tr>
<tr>
<td>log-likelihood</td>
<td>-27352.78</td>
<td>-27351.48</td>
<td>-27345.05</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. Statistical significance: ***p<0.01; **p<0.05; *p<0.1.
<table>
<thead>
<tr>
<th></th>
<th>Cutoff dist. = 0.5 mi</th>
<th></th>
<th>Cutoff dist. = 0.25 mi</th>
<th></th>
<th>Cutoff dist. = 0.1 mi</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NPL-1</td>
<td>NPL-2</td>
<td>NPL-1</td>
<td>NPL-2</td>
<td>NPL-1</td>
<td>NPL-2</td>
</tr>
<tr>
<td>delinquency spillover effect</td>
<td>1.1210*</td>
<td>2.3226***</td>
<td>0.7677</td>
<td>2.1307***</td>
<td>0.3796</td>
<td>2.1194***</td>
</tr>
<tr>
<td></td>
<td>(0.6459)</td>
<td>(0.4629)</td>
<td>(0.5698)</td>
<td>(0.3942)</td>
<td>(0.4732)</td>
<td>(0.3217)</td>
</tr>
<tr>
<td>foreclosure contagion effect</td>
<td>0.0058***</td>
<td>0.0048***</td>
<td>0.0058***</td>
<td>0.0044***</td>
<td>0.0059***</td>
<td>0.0037***</td>
</tr>
<tr>
<td></td>
<td>(0.0012)</td>
<td>(0.0011)</td>
<td>(0.0013)</td>
<td>(0.0011)</td>
<td>(0.0013)</td>
<td>(0.0010)</td>
</tr>
<tr>
<td>owner</td>
<td>0.0585**</td>
<td>0.0593**</td>
<td>0.0585**</td>
<td>0.0594**</td>
<td>0.0584**</td>
<td>0.0591**</td>
</tr>
<tr>
<td></td>
<td>(0.0270)</td>
<td>(0.0270)</td>
<td>(0.0270)</td>
<td>(0.0269)</td>
<td>(0.0270)</td>
<td>(0.0268)</td>
</tr>
<tr>
<td>square footage</td>
<td>-0.3084***</td>
<td>-0.3061***</td>
<td>-0.3088***</td>
<td>-0.3054***</td>
<td>-0.3088***</td>
<td>-0.2998***</td>
</tr>
<tr>
<td></td>
<td>(0.0266)</td>
<td>(0.0257)</td>
<td>(0.0266)</td>
<td>(0.0253)</td>
<td>(0.0268)</td>
<td>(0.0241)</td>
</tr>
<tr>
<td>bedrooms</td>
<td>0.0261</td>
<td>0.0246</td>
<td>0.0263</td>
<td>0.0245</td>
<td>0.0263</td>
<td>0.0233</td>
</tr>
<tr>
<td></td>
<td>(0.0194)</td>
<td>(0.0189)</td>
<td>(0.0194)</td>
<td>(0.0187)</td>
<td>(0.0195)</td>
<td>(0.0180)</td>
</tr>
<tr>
<td>log property value</td>
<td>0.8149***</td>
<td>0.8062***</td>
<td>0.8157***</td>
<td>0.8033***</td>
<td>0.8169***</td>
<td>0.7930***</td>
</tr>
<tr>
<td></td>
<td>(0.0325)</td>
<td>(0.0325)</td>
<td>(0.0326)</td>
<td>(0.0325)</td>
<td>(0.0327)</td>
<td>(0.0328)</td>
</tr>
<tr>
<td>LTV_60to80</td>
<td>0.4499***</td>
<td>0.4487***</td>
<td>0.4499***</td>
<td>0.4485***</td>
<td>0.4498***</td>
<td>0.4472***</td>
</tr>
<tr>
<td></td>
<td>(0.0495)</td>
<td>(0.0494)</td>
<td>(0.0495)</td>
<td>(0.0494)</td>
<td>(0.0495)</td>
<td>(0.0494)</td>
</tr>
<tr>
<td>LTV_80to100</td>
<td>0.8103***</td>
<td>0.8080***</td>
<td>0.8104***</td>
<td>0.8078***</td>
<td>0.8104***</td>
<td>0.8058***</td>
</tr>
<tr>
<td></td>
<td>(0.0479)</td>
<td>(0.0479)</td>
<td>(0.0479)</td>
<td>(0.0479)</td>
<td>(0.0479)</td>
<td>(0.0478)</td>
</tr>
<tr>
<td>LTV_gt100</td>
<td>0.6659***</td>
<td>0.6630***</td>
<td>0.6661***</td>
<td>0.6626***</td>
<td>0.6664***</td>
<td>0.6605***</td>
</tr>
<tr>
<td></td>
<td>(0.0617)</td>
<td>(0.0617)</td>
<td>(0.0617)</td>
<td>(0.0617)</td>
<td>(0.0617)</td>
<td>(0.0616)</td>
</tr>
<tr>
<td>block group dummies</td>
<td>included</td>
<td>included</td>
<td>included</td>
<td>included</td>
<td>included</td>
<td>included</td>
</tr>
<tr>
<td>log-likelihood</td>
<td>-27351.48</td>
<td>-27345.05</td>
<td>-27351.97</td>
<td>-27344.11</td>
<td>-27352.47</td>
<td>-27341.00</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. Statistical significance: ***p<0.01; **p<0.05; *p<0.1.
3.3 Counterfactual Studies

To illustrate the policy relevance of our empirical model and estimation strategy, we carry out the following two counterfactual studies. In the first study, we hypothetically remove properties in foreclosure, one at a time, from the data, and calculate the corresponding reduction in the aggregate delinquency level. More specifically, we first calculate the predicted delinquency probability for every borrower in the study region and add the probabilities up to obtain the initial aggregate delinquency level. Then, we remove a foreclosure from the study region and re-calculate the predicted delinquency probability for every borrower to get the new aggregate delinquency level. Taking the difference between the two aggregate delinquency levels gives the reduction in the aggregate delinquency level from removing that foreclosure. We then repeat this exercise for every foreclosure in the study region to obtain the corresponding reduction in the aggregate delinquency level. Table 5 reports the summary statistics of the reductions based on the logit, NPL-1 and NPL-2 estimates in Table 3. From the table, we can see that the marginal effect of removing a property in foreclosure tends to be understated when the spillover effect is ignored (logit) or inconsistently estimated due to the missing data problem (NPL-1). This exercise sheds light on the importance of correctly estimating the delinquency spillover effect in evaluating the effectiveness of a foreclosure prevention program.

Table 5: Aggregate Delinquency Reduction from the Removal of a Neighboring Foreclosure

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>0.16</td>
<td>0.09</td>
<td>0</td>
<td>0.63</td>
</tr>
<tr>
<td>NPL-1</td>
<td>0.19</td>
<td>0.12</td>
<td>0</td>
<td>0.81</td>
</tr>
<tr>
<td>NPL-2</td>
<td>0.22</td>
<td>0.15</td>
<td>0</td>
<td>1.14</td>
</tr>
</tbody>
</table>

In the second study, we add a constant $c$, which can be interpreted as a mortgage payment reduction, to the utility function (1) of all mortgage borrowers in the study region. The dotted, dashed, and solid lines in Figure 5 represent, respectively, the predicted percentage reduction in delinquency rates as $c$ increases, based on the logit, NPL-1 and NPL-2 estimates.
in Table 3. Similar to the first study, we can see that the marginal effect of loan payment reduction is understated when the spillover effect is ignore (logit) or inconsistently estimated due to the missing data problem (NPL-1).

4 Conclusion

This paper studies the spillover effect of home mortgage delinquencies using a discrete-choice spatial network model. To address the missing data issue in the dependent variable of the model, we propose a modified NPL algorithm that utilizes both sampled loan performance data and public records of all residents in the study region. We carry out Monte Carlo simulations to show that the proposed estimator works well in finite samples and ignoring
this missing data issue leads to a downward bias of the estimated spillover effect. We provide an empirical illustration of our method and conduct some counterfactual experiments to demonstrate the importance of correctly estimating the spillover effect.

Although the motivation of this paper comes from the missing data problem in the loan performance data, the applicability of the proposed method is not limited to this specific setting. As the econometric model described in Section 2.1 is very general, this method can be easily extended to address the missing data issue in the dependent variable of various discrete-choice network models.

References


Griffith, A. (2020), Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. Working paper, University of Washington.


