Data Shop

Data Shop, a department of Cityscape, presents short articles or notes on the uses of data in housing and urban research. Through this department, the Office of Policy Development and Research introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to david.a.vandenbroucke@hud.gov for consideration.

Tracking Individuals Pre- and Post-Foreclosure

Christos Makridis
Massachusetts Institute of Technology (MIT) Sloan School of Management

Michael Ohlrogge
New York University School of Law

Disclaimer
All opinions expressed here are strictly those of the authors and do not reflect the views of affiliated organizations or the U.S. Government.

Abstract

In this article, the authors examine how it is possible to construct records that track 1.4 million households that experience foreclosure from their pre-foreclosure to post-foreclosure residences. These records were created by merging two powerful sets of data: county registrar of deeds records (licensed from CoreLogic, Inc.) and consumer mail marketing data (licensed from RefUSA). The article starts with a description of the county registrar of deeds data and how it can be used to create a dataset of mortgages and outcomes (including foreclosure). The authors proceed to describe the nature of the consumer marketing data from RefUSA, how it can be used to track households as they move locations, and then how that data can be merged with the mortgage records constructed from registrar of deeds data. The article also includes discussions of how those combined records can be merged with the Home Mortgage Disclosure Act (HMDA) database, using GIS software, to learn additional demographic information (income, race, and so on) about individuals with mortgages and foreclosures

Introduction

Residential mortgages and their foreclosures played a pivotal role in the recent financial crisis. A host of high-quality data sets exist with millions, billions, even trillions of observations on residential mortgage origination, performance, delinquency, and foreclosure. Those data sets have given rise to hundreds, if not thousands, of academic papers investigating which individuals got residential mortgages, on what terms, how those mortgages performed, and the implications of those findings for the broader economy. For all of this richness in investigation though, data limitations have made it difficult or impossible for researchers to accurately track where specific individuals move after foreclosure.

More than 6 million households and 10 million individuals experienced mortgage foreclosures from 2007 to 2013 (CoreLogic, Inc., 2017). Thus, understanding what happened to those people is important from macro- and microeconomic perspectives and for documenting the human impacts of one of the most significant financial events of the postwar period in the United States. In our companion article (Makridis and Ohlrogge, 2018), we take on one such effort to improve understanding of post-foreclosure outcomes by examining whether individuals move to areas with better or worse labor market opportunities after foreclosures and what predicts the significant heterogeneity among individuals in those outcomes.

In this article, we present details about our new method for constructing data that tracks more than 1.4 million households that experienced mortgage foreclosure, identifying the precise addresses at which they lived before and after the end of their mortgages. Our approach proceeds through several steps.

First, we use parcel-level data, licensed from CoreLogic, Inc. and gathered from county registrars of deeds to identify the precise addresses and dates for which ownership of a piece of land changes hands due to a completed foreclosure action. Second, we use the street addresses (house number, unit number [if applicable], street name, city, state, and ZIP Code) on these foreclosure records to match them to a consumer mail marketing database licensed from RefUSA. Third, we use the unique household identifiers obtained from RefUSA to track individuals to their next address after a foreclosure. Finally, to obtain additional information about the individuals who experience foreclosure (such as their income at time of mortgage origination, race, and so forth), we match the foreclosure records with data from the Home Mortgage Disclosure Act (HMDA).

To our knowledge, this is the first and only data set to be used in academic research that identifies individuals who have lost their houses to foreclosure and then shows where they move after that foreclosure. Furthermore, although our method depends on two types of data that must be licensed for a fee, such licensing is available to any researcher with the requisite budget, and for both types of data, multiple sources are available from which to license the data, thus increasing the options, availability, and affordability of building such data.

The remainder of this article proceeds as follows. In the next section, we review existing mortgage data sets and their limitations with respect to identifying outcomes of individuals post-foreclosure. We then describe how we constructed records of mortgage originations and outcomes from county registrar of deeds data. In the following section, we describe how we merged this mortgage data with the RefUSA mail marketing database to identify where individuals move post-foreclosure. In the next section, we discuss how we obtained additional information about the individuals subject to foreclosure by merging our mortgage records with the HMDA database. In the final section, we discuss alternative vendors for the two key types of proprietary data we use in our data construction.

Limitations of Existing Mortgage Data Sets

The reason that studying the post-foreclosure outcomes for households has been difficult lies in the nature of how most mortgage data sets used for academic research are generated. Many mortgage data sets, both public and proprietary, are ultimately generated by mortgage servicers and distributed to mortgage investors.[1] After a mortgage foreclosure occurs, those who service or own mortgages have relatively little reason to track where individuals move.[2]

Another type of mortgage data set comes from government records of mortgage originations. The HMDA data set and the registrar of deeds data (described further herein) are examples; however, HMDA does not track any information past a mortgage's origination. Deed registries generally provide the names of individuals who take out mortgages and own land. Then in theory it might be possible to search such data by the names of individuals to discover where they move after foreclosure. First-and-last name combinations are far from unique across the country; in any case, those records will be a small and unrepresentative subsample of the population who are able to buy property again immediately following a mortgage foreclosure.

The only other data set similar to ours for tracking individuals post-foreclosure is that created by the Federal Reserve Bank of New York (FRBNY), using Equifax credit data (see Brevoort and Cooper, 2013; Lee and Van der Klaauw, 2010; Molloy and Shan, 2013). In addition to this data being proprietary to Federal Reserve researchers, it has three significant limitations compared with the data we construct. First, this data set identifies only individuals who have begun the foreclosure process. Given that roughly one-half of foreclosure starts never result in a completed foreclosure (see, for example, Federal Housing Finance Agency, 2018), this introduces a substantial amount of uncertainty to the data. Second, the FRBNY-Equifax data set identifies an individual's location only when a credit report is requested for that person; thus, if an individual moves post-foreclosure but then does not have any credit reports run, that person will seem to have not moved at all. Finally, we track roughly 1.4 million households that experienced foreclosure between 2006 and 2011, compared with only 330,000 households that experienced foreclosure between 2000 and 2009, as tracked in the FRBNY-Equifax data (Brevoort and Cooper, 2013).[3]

[1] Publicly available sets such as these include those from Fannie Mae, Freddie Mac, and the Columbia Collateral File. Private ones that must be licensed for a fee include those from firms such as BlackBox and CoreLogic, Inc.

[2] In particular, in a large number of states, antideficiency statutes explicitly prohibit mortgage owners from trying to collect from borrowers any amounts that remain unpaid on a mortgage loan after foreclosure of a house backing that loan. Even where those statutes do not apply, attempting to collect money beyond the value of the housing collateral from a foreclosed borrower is frequently not economically practicable.

[3] That 330,000 figure is only for foreclosure starts, thus representing perhaps one-half that number of completed foreclosures.

Building Mortgage Records from Registrar of Deeds Data

Throughout the United States, property law generally requires that an interest in land, such as an ownership or mortgage interest, be registered with the county in which the property is located if that interest is to be legally enforceable. Those registries are then made publicly available by counties so that, for instance, a person purchasing property can be assured that the seller has the legal right to dispose of that asset. Commercial companies, such as CoreLogic, Inc. and its predecessors, have in turn gathered this information from the roughly 3,000 U.S. counties, cleaned and standardized it, and licensed it for academic and other research purposes.

County registrars of deeds create unique parcel identification numbers by which records on individual properties can be traced over time.[4] We thus start by identifying in our data when each new loan is taken out on a property, whether that loan is for the initial purchase of the property, a refinance of a purchase loan, or a home equity line of credit. For each such loan, we then observe the set of subsequent records tied to the property's unique identifier.

If, for instance, we observe an initial purchase loan taken out on a property and no subsequent records on that property until 6 years later, at which point we observe a new (non-foreclosure) sale of that property from the prior owners to new owners, then we conclude that the initial loan ended in a prepayment. If a loan is a purchase loan or a refinance of a purchase loan, and we observe a subsequent loan of comparable size (adjusted for amortization) being taken out on the same property, with no interceding changes of that property's ownership, we likewise conclude that the loan was refinanced and thus ended in a prepayment. If, however, we observe a loan registered on a property, and then we observe the property next changing hands through a foreclosure transaction—with no intervening prepayment events—then we conclude that the given loan ended in a foreclosure.

As a practical matter, if our only objective were to track people from before and after foreclosure, our procedure here could be simpler. In particular, we would need to identify only the dates and addresses at which foreclosures occur and then match those addresses to the unique family IDs assigned to them by the RefUSA data, as described later in this article. We also, however, wish to match those loan records to the HDMA data to gain more information about the people whose homes are foreclosed on. We must therefore generate the more precise information on the loans that are being foreclosed on, such as what type of loan (purchase, refinance, home equity, and so on) it was, when the loan was originated, who the lender was, and so forth.

Out of a starting set of roughly 500 million property-level records in the deed data we license, we derive approximately 96 million unique mortgage records through these procedures. In general, multiple property-level records are associated with each mortgage (for instance, one or sometimes

[4] Some of the work of companies such as CoreLogic, Inc. also involves ensuring the uniqueness and continuity of these identifiers.

multiple records when the mortgage is taken out as well as when it is extinguished). We also exclude records on commercial properties, loans for construction and development of subdivisions and large groupings of residential properties, and other records not directly relevant for tracking residential mortgages to individual homeowners.[5]

Matching Deed Mortgage Records to Consumer Marketing Data

Having created mortgage records and identified individuals whose properties were foreclosed on, we next draw on the consumer mail marketing database licensed from RefUSA. RefUSA's primary clients are companies or charitable organizations that target individuals with advertisements and solicitations. The foundation for RefUSA's database comes from proprietary data they license from the U.S. Postal Service. That database incorporates information from individuals who fill out official "change of address" forms when they move, as well as a variety of other government forms through which individuals disclose new addresses. RefUSA then supplements that data through partnerships with credit card companies, magazines, utilities, and similar entities to which people voluntarily report changes in address. Any person who has moved and has subsequently seen catalogs and charitable solicitations follow them, unprompted, to their new address will be familiar with the ability of firms like RefUSA to track individuals, even without any deliberate actions by those individuals to disseminate their new addresses. RefUSA has compiled this information into a yearly panel of roughly 130 million U.S. households.

RefUSA assigns each household a unique identification number to track those households as they move to new locations. Each yearly entry in the panel includes a full street address. Thus, one can readily identify households that move by filtering for those that change street addresses from one year to the next. The specific data we license from RefUSA spans years 2006 to 2012; thus, we are able to identify households that move in years 2006 through 2011.

We match moving households in RefUSA with foreclosed households from the deed data on the basis of the exact correspondence of house number (which includes unit number, if applicable), street name, state, ZIP Code, and year of move or foreclosure. If a foreclosure occurs, for instance, in 2009, then depending on when a household's address is updated in the RefUSA data, they may not be recorded as moving until 2010. Thus, we allow a margin of 1 year when matching on the year of move or foreclosure.

In the mortgage data constructed from deed records, we identify 4,088,248 foreclosures occurring between 2006 and 2011. Of those, we are able to match 1,415,241 to records of households shown as moving in the RefUSA data, giving us a match rate of 34.6 percent. Our best understanding of the reason for this low match rate is that not all of the foreclosed households were initially in the

[5] Overall, this construction of mortgage histories out of these public record filings is similar to the approach taken by Ferreira and Gyourko (2015). See also Diamond and McQuade (2016) and Ohlrogge (2018) for additional uses of such public records deed data.

RefUSA database, and for those that were, RefUSA may have experienced delays before identifying and reporting them as having moved.[6]

A natural concern with a low match rate such as this is that the sample of foreclosures for which we identify where the household moved post-foreclosure may not be representative of the full set of foreclosures. We perform several tests to investigate whether this is likely a concern.[7]

First, we consider the initial mortgage amount (in dollars) for foreclosures that we match to the RefUSA moving records and for those that we did not. We construct a quantile-quantile (Q-Q) plot[8] comparing the quantiles of the two distributions. Exhibit 1 presents the results, showing nearly identical distributions between the two sets of foreclosures.[9] Thus, no systematic bias seems to exist, whereby households with smaller mortgages (presumably, with lower incomes) are less likely to be tracked to their post-foreclosure locations.

Next, we investigate whether geographic differences might be present, in which some areas are more likely to have individuals whom we are able to successfully track via RefUSA. We compute the fraction of total observations that each U.S. ZIP Code accounts for in the matched and total sample, finding a correlation between the two samples of 0.92. Although a higher match rate would undoubtedly be desirable, we are nevertheless able to identify a large number (1.4 million) of foreclosures for which we can track the post-foreclosure residence of the household, and no systematic differences seem to be present in either the individual or the geographic characteristics of the individuals for whom we are and are not able to match.

[6] In Makridis and Ohlrogge (2017), we examined another data product from RefUSA that advertised information on the number of people employed at each U.S. business, with updates on a yearly basis. In that paper, we demonstrated that those employment counts are subject to substantial interpolation that may limit the number of applications for which they can be used. By contrast, determining the address of an individual is far easier than determining the number of people employed by a business. As such, we are less concerned about data errors in RefUSA's consumer marketing database. Although such errors undoubtedly exist and unquestionably account for some portion of our failure to match foreclosure records, the tests we describe herein give us some confidence that match failures do not substantially reduce the usefulness of the data.

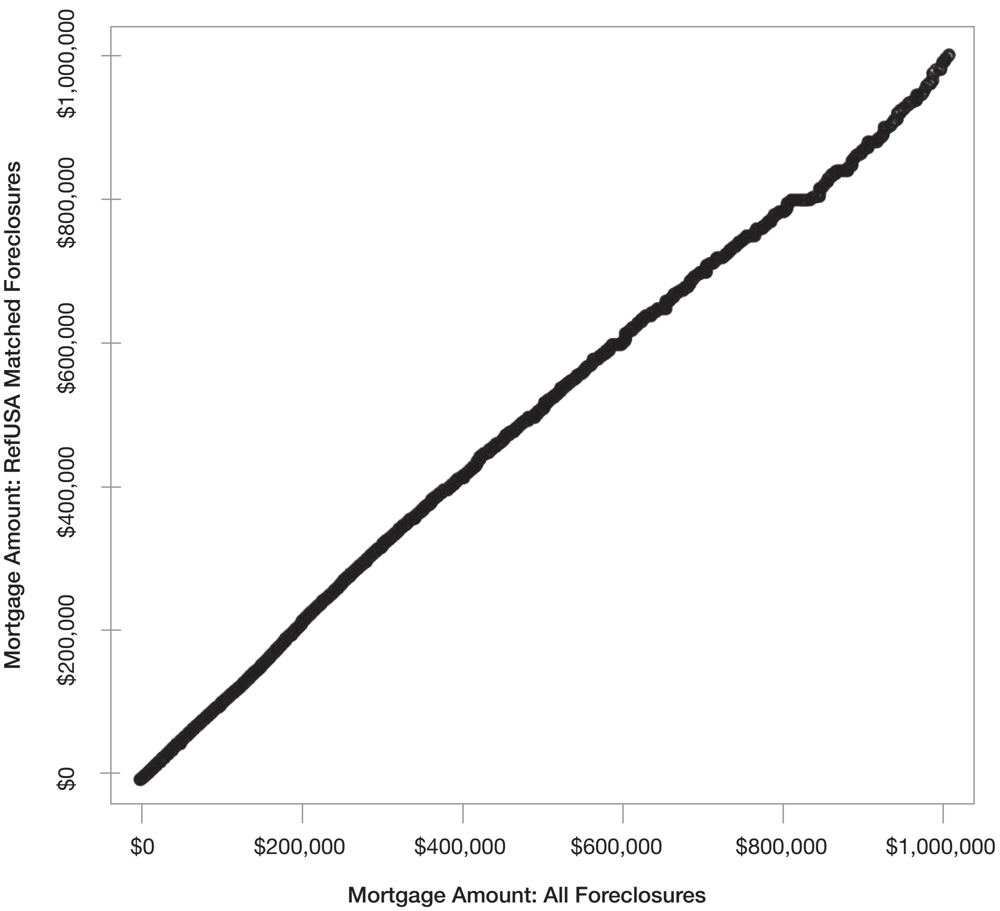
[7] See also Makridis and Ohlrogge (2018) for additional validation tests and discussion.

[8] A Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles (for example deciles, quintiles, quartiles, medians) against each other.

[9] In this plot, we restricted to mortgages less than $1 million in initial value. Only very sparse coverage exists for mortgages higher than that value, making comparisons of distributions much less meaningful.

Exhibit 1

Q–Q Plot—Mortgage Amount All Foreclosures vs. RefUSA Matched Foreclosures

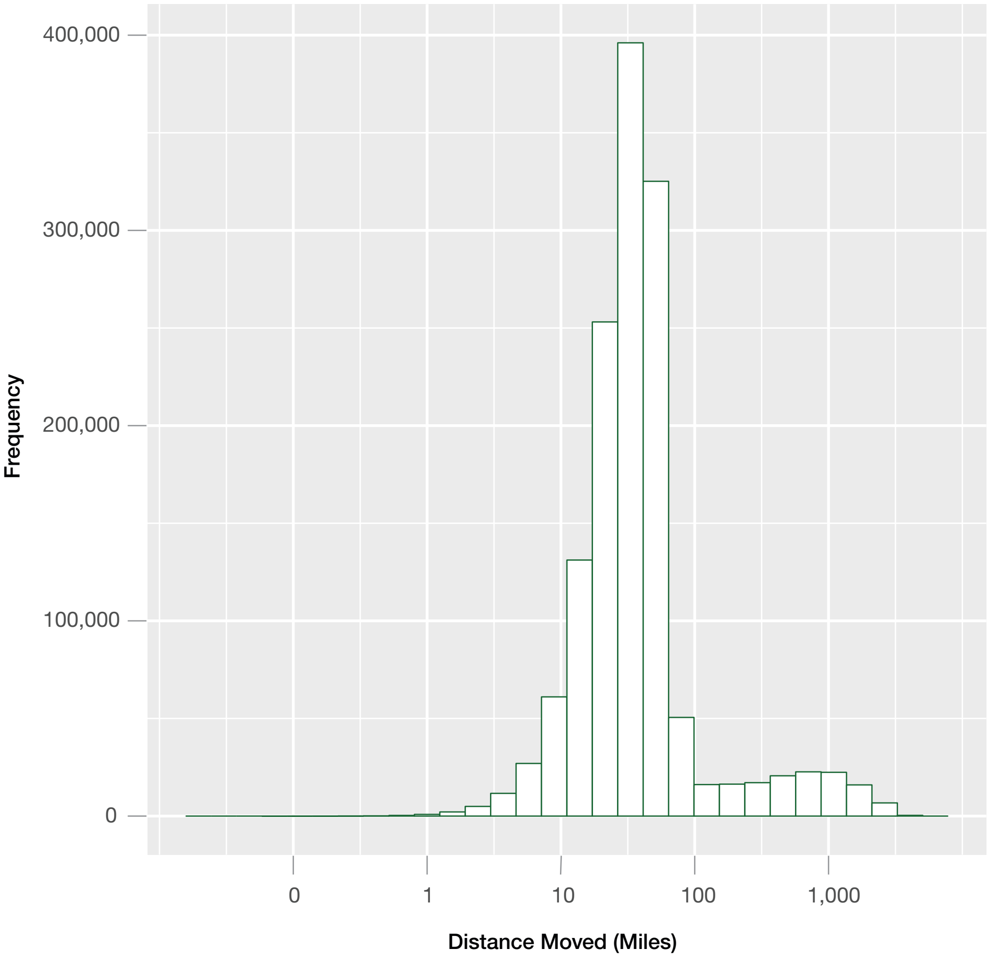


Notes: This plot compares the distribution of the total mortgage amount among the full set of foreclosure records we constructed from the registrar of deeds data and the 34.6 percent of those records for which we were able to identify where the foreclosed household moved afterwards, based on matching the foreclosure records with the RefUSA consumer mail marketing database. If, for instance, more affluent individuals were more likely to be matched, that would be reflected in this plot. We view this Q-Q plot as a sufficient statistic for evaluating comparability of the matched and unmatched samples along the dimension of mortgage amount.

On the basis of those results, which suggest that the data reliability is still preserved given the matching process, we now present several summary results that demonstrate some of the investigations this data set makes possible. First, exhibit 2 presents the distribution of distances moved by people post-foreclosure. The exhibit shows that the large majority moved between 10 and 100 miles to a new location, with smaller numbers moving lesser and greater distances. Exhibit 3 presents specific empirical quantiles for the distribution of distances moved. The median distance moved is 34 miles—comparable, for instance, to a move from Oakland, California to San Jose, California. The 90th percentile for distance moved was 95 miles—comparable to a move from Philadelphia, Pennsylvania to New York, New York. The 95th percentile for distance moved was 548.1 miles—comparable to a move from Memphis, Tennessee to Atlanta, Georgia.

Exhibit 2

Distance Moved After Foreclosures



Notes: This graph depicts the distance (measured in miles) that individuals moved after foreclosure. As evidenced here, the vast majority of people moved within 10 and 100 miles, with smaller numbers moving less than 10 miles and a moderate number moving between 100 and 1,000 (or more) miles to a new location post-foreclosure. See exhibit 3 for specific quantiles of the distribution presented here.

Exhibit 3

Distance Moved After Foreclosure—Percentiles

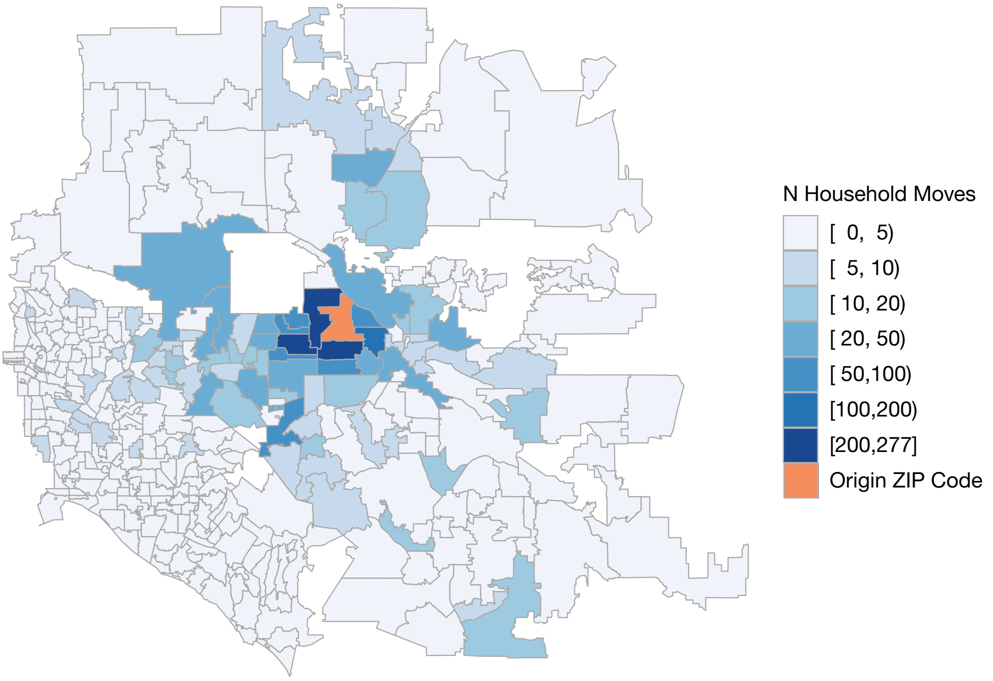
Quantile (%)	1	5	10	25	50	75	90	95	99
Miles Moved	3.8	8.8	12.7	21.5	34.0	49.1	95.0	548.1	1,744.9

Notes: This table presents several of the empirical percentiles for distance moved after foreclosure, corresponding to the distribution plotted in the histogram in exhibit 2.

Next, exhibit 4 focuses on one of the ZIP Codes (92336) in San Bernardino, California that was at the epicenter of the foreclosure crisis. The figure plots the diaspora of people moving from ZIP Code 92336 to a neighboring area.

Exhibit 4

Foreclosure Diaspora from ZIP Code 92336



Notes: This exhibit presents ZIP Code 92336 (depicted here in orange), in San Bernardino, California, which is one of the ZIP Codes with the largest number of foreclosures in the United States. We observe 4,112 foreclosures in this ZIP code from 2006 through 2011 for which we can identify where the household moved after foreclosure. Of that number, 29.3 percent (1,205 households) remained in the given ZIP Code. Of the 2,907 that left the ZIP Code, 2,403 remained within a roughly 25-mile radius of the ZIP Code. This plot depicts how many individuals moved from ZIP Code 92336 to each other ZIP Code within that 25-mile radius.

Matching Deed Mortgage Records to HMDA

To gain additional information on foreclosed households, such as the income reported at the time of taking out the mortgage as well as race and gender information on the borrowers, we also match the mortgage records created from the registrars of deeds data with the HMDA data.[10] Both sets of data are extremely detailed, which makes accurate matching possible, although not without some substantial challenges that must first be overcome. We match on the basis of agreement of census tract, loan amount, year of loan origination, lender name, whether the loan is a purchase or a refinance loan, and whether the loan is a conventional mortgage, FHA backed, or VA backed.

Census tracts are extremely small pieces of land, making them ideal for such a matching operation. The U.S. Census Bureau targets between 2,000 and 8,000 people in each tract and re-adjusts their boundaries after each decennial census to maintain their populations within that range. As supplied by CoreLogic, Inc., the deed data comes with the census tracts as designated by the 2010 census. HMDA data reports 2010 census tracts starting in the year 2013.[11] For loans originated from 2003 to 2012, HMDA data uses the 2000 census tracts, and for loans from 1993 to 2002, HMDA data uses the 1990 census tracts. The boundaries of census tracts can change over time. Thus, for loans in the deed data originated before 2013, it is necessary to first compute the 2000 or 1990 census tract (depending on the loan's origination date) before matching with HMDA data.

We use ArcGIS software and geographic shape files designating the boundaries of the census tracts to map the addresses in the deed records to census tracts from the appropriate census year.[12] The shape files we use in this process are from the National Historical Geographic Information System (NHGIS).

A final challenge is that the names of lenders in the two databases are not precisely the same (for instance, one may contain "Bank of America," whereas the other lists "Bank of America, NA"). We therefore use an algorithm based on the Levenshtein distance between strings to pair lender names that are not exact matches.

In total, we are able to match 68 percent of the mortgages in the deed data to HMDA data, for a total of 65 million mortgage records, including both foreclosed and non-foreclosed mortgages.[13] In Makridis and Ohlrogge (2018), we perform analyses similar to those discussed previously to validate that the distributions of observable characteristics were very similar in the set of foreclosures that we did and did not successfully match to HMDA data.

[10] For 2006, we obtained HMDA data from the Inter-university Consortium for Political and Social Research: https://doi.org/10.3886/ ICPSR24612.v2; for later years, we obtained it directly from the Federal Financial Institutions Examination Council: https://www.ffiec.gov/hmda/hmdaproducts.htm; for earlier years, we obtained CDs from the National Archives.

[11] For details on HMDA data reporting, see, for example, https://www.ffiec.gov/hmda/guide.htm.

[12] In particular, we first found the latitude and longitude coordinates for each address. We then matched those coordinates to shape files for the census tracts.

[13] This success rate closely matches that of other researchers who have performed similar such matching. See, for example, work by Nancy Wallace: https://bfi.uchicago.edu/sites/default/files/file_uploads/3_Wallace_MFM.pdf. In calculating the percentage of deed records matched to HMDA data, we first removed certain deed records that would obviously not be in HMDA data, such as those for loans made to develop large tracks of land for new subdivisions, loans for commercial properties, and so forth.

Alternative Data Sources

The comprehensive country registrar of deeds data that we use in this study was licensed for a fee from CoreLogic, Inc.; however, other companies also sell comparable data, offering researchers the potential to identify the best vendor for their purposes and to potentially negotiate better prices. Attom Data Solutions[14] is one such option. Some U.S. counties may well also provide this information for free in a form accessible to researchers, although we have not specifically investigated that possibility. Similarly, although we use consumer mail marketing data licensed from RefUSA, other firms, such as the Experian credit rating agency, also offer similar products for license.[15]

Acknowledgments

We would like to thank Brian Wang for diligent research assistance and Kris Kasianovitz for assistance in acquiring the data used in this project. Michael thanks the Olin Foundation for Law and Economics and the Stanford Advanced Financial Technologies (AFT) Lab for financial support. Christos thanks the National Science Foundation Graduate Research Fellowship and the Stanford Institute for Economic Policy Research Shultz Fellowship for financial support. All opinions expressed here are strictly our own and not those of any institution with which we are affiliated. Michael Ohlrogge was a Ph.D. candidate in Stanford University's Department of Management Science and Engineering at the time this article was written.

Authors

Michael Ohlrogge is an assistant professor at New York University School of Law.

Christos Makridis currently serves on the White House Council of Economic Advisers, is a Digital Fellow at the MIT Sloan School of Management's Initiative on the Digital Economy, a non-resident fellow at the Harvard Kennedy School of Government's Cyber Project, and Baylor University's Institute of Religious Studies.

References

Brevoort, Kenneth P., and Cheryl R. Cooper. 2013. "Foreclosure's Wake: The Credit Experiences of Individuals Following Foreclosure," *Real Estate Economics* 41 (4): 747–792. DOI: 10.1111/1540-6229.12021.

CoreLogic, Inc. 2017. *United States Residential Foreclosure Crisis: Ten Years Later*. Irvine, CA: CoreLogic, Inc. https://www.corelogic.com/research/foreclosure-report/national-foreclosure-report-10-year.pdf.

[14] The website for Attom Data Solutions is https://www.attomdata.com/data/property-data/recorder-data/.

[15] Experian's website is http://www.experian.com/small-business/mailing-lists.jsp.

Diamond, Rebecca, and Timothy McQuade. 2016. Who Wants Affordable Housing in their Backyard? An Equilibrium Analysis of Low Income Property Development. NBER Working Paper No. w22204. Cambridge, MA: National Bureau of Economic Research.

Federal Housing Finance Agency (FHFA). 2018. *Foreclosure Prevention Report*. Washington, DC: Federal Housing Finance Agency. https://www.fhfa.gov/AboutUs/Reports/ReportDocuments/FPR_1Q2018.pdf.

Ferreira, Fernando, and Joseph Gyourko. 2015. A New Look at the U.S. Foreclosure Crisis: Panel Data Evidence of Prime and Subprime Borrowers from 1997 to 2012. NBER Working Paper No. 21261. Cambridge, MA: National Bureau of Economic Research.

Lee, Donghoon, and Wilbert Van der Klaauw. 2010. *An Introduction to the FRBNY Consumer Credit Panel*, No. 479. New York: Federal Reserve Bank of New York.

Makridis, Christos, and Michael Ohlrogge. 2018. Moving to Opportunity? The Geography of the Foreclosure Crisis & the Importance of Location. Working paper (April 14). Stanford, CA: Stanford University; Cambridge, MA: Massachusetts Institute of Technology. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3162905.

———. 2017. "Validating RefUSA Micro-Data with the Longitudinal Employer-Household Dynamics Data," *Economics Letters* 152: 83–87. DOI: 10.1016/j.econlet.2017.01.001.

Molloy, Raven, and Hui Shan. 2013. "The Post-Foreclosure Experience of U.S. Households," *Real Estate Economics* 41 (2): 225–254. DOI: 10.1111/j.1540-6229.2012.00344x.

Ohlrogge, Michael. 2018. Bank Capital and Risk Taking: A Loan Level Analysis. Working paper. Stanford, CA: Stanford University, Stanford Center for Financial and Risk Analytics. https://papers.ssrn.com/sol3/papers.cfm?abstract_id= 3053240.