Jasper Evans, Krehl Kasayan, Paul Kiefer, Pablo Suarez
INST 737
Milestone 1 Report
2/25/2024

[GitHub Page Link]

Research Question:

Is it possible to forecast the per capita rate of foreclosure for a census tract in Prince George's County, Maryland, using the following independent variables:

1. The median age of tract residents as recorded by the 2016-2020 ACS.
2. The percentage of the tract's population identified as non-Hispanic white during the 2016-2020 ACS.
3. The change in the percentage of the tract's population identified as non-Hispanic white between the 2006-2010 ACS and 2016-2020 ACS.
4. Tract-level poverty rate as of the 2016-2020 ACS.
5. The change in reported median income household income in the tract between the 2006-2010 ACS, 2011-2015 ACS, and 2016-2020 ACS.
6. The change in the ratio of mortgaged housing units to total housing units per tract recorded by the 2006-2010 ACS, 2011-2015 ACS, and 2016-2020 ACS.
7. The change in the ratio of owner-occupied housing units to total housing units per tract recorded by the 2006-2010 ACS, 2011-2015 ACS, and 2016-2020 ACS.
8. The share of all units in a tract built before and after 2009 as recorded in the 2016-2020 ACS.
9. The mean number of bedrooms per house as recorded in the 2016-2020 ACS.
10. The median reported value of homes as recorded in the 2016-2020 ACS.

Foreclosure rates surged across Maryland after the state's pandemic-era foreclosure moratorium ended in June 2021. Prince George's County residents were hit far harder than their neighbors in Montgomery County.

When a property owner – often a homeowner – is at risk of falling behind on mortgage payments, early intervention can prevent a foreclosure and the associated externalities, including impacts on the borrower's credit.

Intervention programs – both privately and publicly funded – are only effective if administrators know where to target their outreach and spending. A predictive model that uses widely-available American Community Survey data and a dataset of foreclosures maintained by the Prince George's County office of planning could be a tool to help target foreclosure mitigation efforts.

Our research question will, we hope, lead us to that tool.

From a technical perspective, the variables available to us through the American Community Survey – including point-in-time measures of median household income and year-over-year

measures of changing demographics – are relatively well-tracked data points that serve as proxies for some of the root causes of foreclosures.

Generational wealth gaps, for instance, leave Black mortgagees at greater risk of foreclosure than white mortgagees, so including the change in the percentage of a census tract's population identifying as non-Hispanic white over a decade is an imperfect proxy for changes in the financial backgrounds of mortgagees in said tract.

Other forces involved in foreclosures, including social pressures that lead homebuyers to purchase a home despite being unable to afford the interest rate of their mortgage, are harder to track through ACS data. However, given that Prince George's County is made up largely of suburban communities that have seen influxes of former city-dwellers seeking relatively affordable homeownership, it may be possible to link increases in the ratio of mortgaged housing units to total housing units in a given census tract with a boom in suburban development.

If there is a correlation between a change in the number of housing units in a given tract or the average number of bedrooms among homes in a given tract, it could signal a relationship between certain types of suburban communities – single-family subdivisions with larger homes, for instance – and risk of foreclosure.

State of the Art

Paper 1: Analyzing foreclosures Among High-Income Black/African American and Hispanic/Latino Borrowers in Prince George's County, Maryland

This study was conducted by researchers from George Mason University, and the National Community Reinvestment Coalition. This study merged data from the US Census Bureau, the Home Mortgage Disclosure Act (HDMA), and Lender Processing Services (LPS) to analyze the likelihood of foreclosure in Prince George's County, Maryland, using a logistic regression model. The study found that borrowers in Black/African American neighborhoods with high income were 42% more likely, and Hispanic/Latino neighborhoods with high income were 159% more likely than borrowers in non-Hispanic White neighborhoods to go into foreclosure.

This study leveraged the following variables from HDMA data: borrower income, race/ethnicity; the following variables from LPS data: race and ethnicity, mortgage characteristics, mortgage payment information; and the following variables from US Census Data at the tract or zip code level: Home Price index, median year structure built, neighborhood race and ethnicity proportions (calculated), number of owner-occupied homes, median household income. FICO scores and mean incomes presented themselves as strong variables for use in the prediction of foreclosure, as higher a proportion of borrowers with low FICO scores (below 640) were affected by foreclosure than were borrowers with medium or high FICO scores (above 720), as well as borrowers above or below specific mean income levels defined within the study.

Paper 2: Housing vacancy and urban growth: explaining changes in long-term vacancy after the US foreclosure crisis

This paper investigates the impact of the Great Recession on long-term vacancies in the USA, focusing on three types of metropolitan areas: Weak-Growth Metros, Hard Hit Metros (Economically Declining), and Strong-Growth Metros. Using multivariate analyses with data from the US Postal Service during the US housing recovery period from 2011 to 2014, the study reveals distinct patterns and factors influencing long-term vacancies in these areas.

In Weak-Growth Metros, larger increases in long-term vacancies were observed in depressed neighborhoods, attributed to a high proportion of African Americans and single-family home renters. Hardhit Metros experienced rising long-term vacancies, particularly in neighborhoods with high poverty levels and more townhomes and condominiums.

Conversely, in Strong-Growth Metros, increases in long-term vacancies were noted in outlying counties. The paper emphasizes the need for researchers and policymakers to recognize the nuanced differences in neighborhood dynamics affecting long-term vacant homes based on the growth trajectory of metropolitan areas, cautioning against generalizations when comparing different regions. A comprehensive understanding of these dynamics is crucial for informed housing policy and research.

Paper 3: Addressing the Foreclosure Crisis: Action Oriented Research in Metropolitan Atlanta

This study was conducted by researchers from Emory University, the Atlanta Regional Commission, the Georgia Institute of Technology, and the Federal Reserve Bank of Atlanta with three primary objectives. The most relevant of the study's objectives to the work Team Two is conducting was to use data analysis to identify neighborhoods in the metropolitan Atlanta area that were most at risk for foreclosure – "hot spots" at the US Census "block" level or tract level and develop a spatial portrait of where/what the hot spots look like.

The study leveraged U.S. Census Data, HDMA data, and monthly foreclosure sales reports to identify hotspot tracts in the Atlanta area with concentrations of subprime lending. Subprime lending entails the granting of a mortgage to a borrower with impaired credit scores; these loans often come with much higher interest rates than non-subprime mortgages to compensate the lender for accepting a borrower with lower credit. The researchers also incorporated data from monthly local mortgage foreclosure filings. Through identifying these hotspots, the research then went about identifying the demographic characteristics of the hotspots. The researchers were able to heatmap foreclosures per square mile in the Atlanta metro area. The researchers had not identified the hotspot neighborhood characteristics but would be able to do so after spatially joining census tract data to the address-level foreclosure data in their possession.

Paper 4: Targeting foreclosure interventions: An analysis of neighborhood characteristics associated with high foreclosure rates in two Minnesota counties (Grover, M. et. al, 2008)

Researchers from the Federal Reserve Bank of Minneapolis and Macalester College examine the statistical association of foreclosure sales with social, economic and housing variables as a means to predict high-foreclosure neighborhoods for foreclosure mitigation programs. They rely on 2002 foreclosure sale data from the two core counties in the Twin Cities metropolitan area – Hennepin and Ramsey Counties, Minnesota – as well as 2000 decennial census data and Home Mortgage Disclosure Act (HDMA) population-level credit score data.

The researchers built a predictive model that used foreclosure sales per capita at the census tract level as a dependent variable and the following as independent variables at the census tract level:

1. Percentage of adults with very low credit scores at the census tract level in 1999.
2. 1996-1999 rate of denial on prime home purchase mortgage applications.
3. 1999 subprime refinancing originations per Census 2000 mortgaged unit.
4. 1999 FHA originations per Census 2000 mortgaged units.
5. 2000 minority share of population.
6. 2000 minority share of homeowners.
7. 1990-2000 change in minority share of population.
8. 1990-2000 change in minority share of homeowners.
9. 2000 share of population over 25 with a college degree.
10. 2000 share of population over 18 with a high school diploma or equivalent.
11. Logarithm of 1999 average household income.
12. 2000 rate of unemployment.
13. 2000 share of home-owning households with a head of household younger than 45 years old.

Of those independent variables, the 1999 share of households with very low credit scores, the 2000 minority share of population and homeowners, and the 2000 share of the population over 18 with a high school diploma or equivalent correlated most strongly with high rates of foreclosure per capita at the zip code level.

The false negatives produced by the researchers' prediction model were generally adjacent to correctly predicted high-risk tracts, and most false positives were also adjacent to high-risk tracts. Researchers did, however, identify some false negatives in wealthy neighborhoods that "may not have been targets for mitigation resources even if the areas had been identified in advance."

Relevance to our Model: Given our storage, time and budget constraints, some of the independent variables used by the researchers are out of our reach – namely those provided by HDMA credit data. We can, however, replicate some of the independent variables that correlated most strongly with high rates of foreclosures per capita at the census tract level. We can also test independent variables that track characteristics of the homes themselves – the year of construction, for instance – to learn more about the relationship between home, homeowner and debt.

Our project is a continuation of this prior research, albeit with limited ability to access some data sources used in said research.

Datasets

We use a dataset of foreclosures by address in Prince George's County as the core of our project. The dataset is maintained by the Prince George's County Planning Department and lists foreclosures by address between 2009 and 2024, including the date upon which notice of the foreclosure was submitted to the county. In its original form — as of February 2024 – it contains 71,676 records, including some duplicates.

The core dataset does not, however, include census tract identification numbers. To match addresses to census tracts, we turned to Geocodio, a geocoding service frequently used by news organizations to geolocate and map addresses. Geocodio offers an API, but the API instructions are convoluted and offer little information about how to fetch census tracts rather than geographic coordinates. Instead, we relied on Geocodio's upload service, which allowed us to upload the consolidated address column of our core dataset and, for a small fee, returned a new dataset with geographic data – including a census tract ID – for nearly all of the 71,676 records.

Either because of inaccuracies in Geocodio's geolocation service or because of data entry errors on the part of the county, many of the census tracts provided for addresses in the dataset are not in Prince George's County. As we will explain later in this report, these inaccuracies required filtering. In total, the geolocated dataset produced by Geocodio included more than 400 unique census tract identification numbers; after filtering, we will work with fewer than 200.

For demographic, housing, and income data, we relied upon American Community Survey data. That data comes in multiple forms, including the U.S. Census Bureau's 2022 DP04: Selected Housing Characteristics ACS data — a short guide to year of construction and home size data from the 2022 survey — and the data available through the US Census API.

We obtained Census API keys to call census variables directly in our code. Those census variables are:

1. The number of housing units per tract as of the 2006-2010, 2011-2015, and 2016-2020 ACS.
2. The number of owner-occupied housing units per tract as of the 2006-2010, 2011-2015, and 2016-2020 ACS.
3. The number of housing units subject to a mortgage or other loan as of the 2006-2010, 2011-2015, and 2016-2020 ACS.
4. The number of people identified as non-Hispanic white by census tract as of the 2006-2010 and 2016-2020 ACS.
5. The number of people ages 5 and older with incomes below the poverty level as of the 2006-2010 and 2016-2020 ACS.
6. The median age of residents at the census tract level as of the 2016-2020 ACS.

7. The median household income at the census tract level as of the 2006-2010, 2011-2015, and 2016-2020 ACS.

Using the ACS census tract population estimates as of the 2016-2020 survey, we calculated the number of foreclosures per 1,000 residents for each census tract in Prince George's County that remained in our dataset after initial filtering.

We had plans to include other datasets, including address-level tax assessment data for Prince George's County maintained by the SDAT (State Department of Assessments and Taxation) and MDP (Maryland Department of Planning), housing inspections violations data maintained by the Prince George's County planning department, and Maryland mortgage application data collected by the Consumer Financial Protection Bureau under the Home Mortgage Disclosure Act.

Unfortunately, the 100 MB cap on file storage in GitHub repositories makes it difficult for us to reference these additional datasets in our code, and the Git Large File Storage system is not straightforward. Additionally, while we could use the core foreclosure dataset in combination with the assessment and inspection violations datasets on an address level, the inconsistencies in address formatting make joining those datasets challenging. Fetching census tract identification numbers for the addresses included in the assessment or inspection violation datasets through Geocodio's service would cost hundreds of dollars — money that we do not have available for this project.

Data Cleaning Efforts

The initial foreclosure dataset includes thousands of duplicate records, including records suggesting that the same address faced foreclosure multiple times in a month. We assume th. After initially joining the core foreclosure dataset with a dataframe that included census tract IDs for each address – a process that, because of the duplicate addresses in the original dataset, creates additional duplicate records – we need to de-duplicate.

To identify duplicate records, we search for rows with matching values in four columns: "propertyid", "city.y", "street_address" and "submitteddate". We use "propertyid", "street_address" and "city.y" for redundancy; records with matching values in those columns and matching values in the "submitteddate" column, which records the date a foreclosure notice was submitted to Prince George's County, are most likely duplicate records. After identifying these duplicates, we subset to keep only the last instance of each set of duplicate records.

There are, however, some records with street addresses that appear to match and "submitteddate" values within days or months of one another. These sets of apparently matching records could either be duplicates or records from multiple foreclosures within the same condominium building. A cursory search of some of the addresses with near-duplicate foreclosure records suggests that many are indeed records from multiple units within a single multifamily building, meaning the address field omitted the unit number. Because we are unable to distinguish between genuine duplicates and false duplicates when all key fields besides "submitteddate" match, we do not target these records for de-duplication.

In addition to de-duplication, we filtered the dataset to exclude the roughly 120 records for which Geocodio did not generate a census tract identification number and the 16 records geolocated to census tracts outside of Maryland.

After grouping the remaining 71,519 records by census tract, we use an initial ACS variable dataframe — in this case, the 2016-2020 ACS census tract population estimates for Maryland — to filter out non-Prince George's County census tracts from our dataset. Because the ACS data frame includes a county name for each census tract, we are able to filter the ACS data to include only Prince George's County tracts and use a left join to filter our core dataset, leaving us with a single data frame that includes the number of foreclosures per tract and the 2016-2020 ACS population estimate for said tract.

Because our dependent variable is the number of foreclosures per 1,000 people at the census tract level, we then filter out any census tracts for which we cannot calculate a per capita foreclosure rate.

The only remaining missing values are found in the column containing mean reported home values as of the 2016-2020 ACS.

By plotting the distribution of foreclosure rates per capita across our remaining census tracts as a histogram, we can check whether our dependent variable is normally distributed and whether we need to address outliers. The resulting histogram roughly resembles a log-normal distribution with no outliers.

Other Software Engineering Efforts

Much of our project relies on data pulled from the past three five-year ACS surveys. To do this, we parsed the census tract identification number from data frames generated with the Census API and used the resulting tract identification numbers to join each variable-specific ACS data frame with our core dataset. This is a repetitive process that requires only minor adjustments from variable to variable.

We also used the U.S. Census Bureau's 2022 DP04: Selected Housing Characteristics ACS data to assist us in identifying the average number of bedrooms per property across census tracts in Prince George's County and an average year range during which properties in the tracts were built. To clean this dataset, we filtered out all values aside from bedrooms and construction years and removed blank values and columns. With the remaining raw count values, we could create new columns to calculate the averages for our two target variables.

Regarding the average number of rooms per census tract, we multiplied the values of each bedroom column by the corresponding number of bedrooms for that column. The original dataset organized bedroom values by property from zero to five bedrooms. Therefore, if there were 54 units in a tract with one bedroom, this would be calculated as 54 x 1. Then, we created a formula to add the results from each column in the tract and divide the total by the number of properties in the tract. We also used the ROUND function to ensure our results would round to

the nearest whole number. Below is the formula for the first census tract in the data: =ROUND((P3*0+Q3*1+R3*2+S3*3+T3*4+U3*5)/O3,0).

The average construction year data followed a similar process. However, this data is less valuable for our purposes because it is organized into year ranges rather than individual years. Regardless, we assigned each range an arbitrary value. Then, following the same formula as above, we multiplied the number of units built within each date range by the values provided in the dataset. Once we acquired the average number, we calculated the corresponding date range using VLOOKUP and designated a number value for each range.

Before realizing the budget and storage necessary to accomplish the task, we began the process of concatenating address components from the property tax assessment data. The resulting full address column would have been uploaded to Geocodio to generate corresponding census tract identification numbers, which we then would have used to group the records by census tract and calculate the mean assessed home value – a more accurate measure than the self-reported values recorded by the ACS. The same process would have been necessary to calculate the total number of home inspection violations per census tract.

We also began similar efforts to group mortgage loan records by census tract; that dataset already includes census tract identification numbers for each address.

Lastly, we requested a dataset of Airbnb listings in Prince George's County from Inside Airbnb, a nonprofit organization that collects and visualizes data on short-term rental listings. If we obtain said dataset, we would need to pass the addresses through the Geocodio service to calculate the ratio of short-term rentals to total housing units at the census tract level – a novel potential independent variable for our model but one that may not come to pass for budget, storage and timing reasons.

Contributions

Research Question: Developed by All.

State of the Art: Jasper Evans, Krehl Kasayan and Paul Kiefer.

Data Engineering: Paul Kiefer and Pablo Suarez.

Code: Paul Kiefer.

Presentation: Pablo Suarez. All members recorded their slides.