Spatial and Temporal Components of Mortgage Default

USING OPEN-SOURCE DATA TO QUANTIFY RISK

"People hate to think about bad things happening so they always underestimate their likelihood."

Jamie Shipley, The Big Short

ABSTRACT

This paper is an exploratory analysis of open source data provided by federal housing- and mortgage-related agencies. The project is introduced with a brief narrative on the housing market collapse of 2007-2009. I then discuss the recent mass reporting of data from Fannie Mae, Freddie Mac, the Federal Housing Finance Agency, the Consumer Finance Protection Bureau, and the Office for Housing and Urban Development. I select variables and build a model for regressing mortgage default rates in over three hundred Metropolitan Statistical Areas (MSAs), progressing in model development to include elements of space and time.

I. Introduction

As the eight-year anniversary of the collapse of Lehman Brothers nears, there is still little consensus as to what cause the housing bubble and ensuing Great Recession that the global economy is just now starting to pull out of. The collapse in housing prices and construction constituted one of the most dramatic such episodes in the history of the U.S. housing industry. This entailed a large decrease in credit standards, an increased use of subprime lending, billions in lost home equity, and a consequent decline in consumption spending. In addition, the post-collapse financial distress decimated lender balance sheets, and the high level of foreclosures appear to have a had negative effect even on the value of neighboring homes whose mortgages were not in risk of default (Rogers and Winters, 2009).

Accordingly, many researchers have attempted to sort out the causes of the run up and crash in housing. There have been numerous factors suspected proposed as being the cause of the crisis, from lax regulation to tax code changes to irrational consumer expectations of rising housing prices. None of these explanations, however, is capable of fully explaining the housing bubble. Neither is this project.

In reality, the housing bubble and subsequent financial crisis were the result of a confluence of macro- and micro-economic phenomena, upon which academics have assigned varying responsibility. This does not mean the crisis can pass without offering lessons which might one day prevent a future housing crash of such magnitude. Silver linings of the economic collapse and subsequent maneuvers made by federal regulatory and housing-related agencies is the open dissemination of an unprecedented amount of data related to housing markets, lending behavior, and default patterns. One of the most alarming aspects of the housing collapse was that barely anyone saw it coming, and those who did either weren't listened to (or decided to quietly profit off it). I hope this project makes a contribution to the data dissemination by putting primary source data into visualizations and models that can be replicated and improved by others interested protecting the health of the U.S. housing market.

II. Data Aggregation

The GSE's

The Federal National Mortgage Association (Fannie Mae) and Federal Home Loan Mortgage Corporation (Freddie Mac) are government sponsored enterprises - financial services corporations created by the United States Congress. Their intended function is to enhance the flow of credit to targeted sectors of the economy and to make those segments of the capital market more efficient and transparent, and to reduce the risk to investors and other suppliers of capital. The desired effect of the GSEs is to enhance the availability and reduce the cost of credit to the targeted borrowing sectors primarily by reducing the risk of capital losses to investors. In order to do this, they buy mortgages on the secondary market, pool them, and sell them as a mortgage-backed security (MBS) to investors on the open market. This secondary mortgage market increases the supply of money available for mortgage lending and increases the money available for new home purchases.

On September 7, 2008, Federal Housing Finance Agency (FHFA) director James B. Lockhart III announced he had put Fannie Mae and Freddie Mac under the conservatorship. The action has been described as one of the most sweeping government interventions in private financial markets in decades. In all, the GSE's went through a nearly $200 billion government bailout during the financial crisis, paid for by the U.S. government. In 2013, Fannie Mae and Freddie Mac began reporting loan-level credit performance data in 2013 at the direction of their new regulator, releasing an unprecedented amount of open source information on loans originated from 2000 to present, updated quarterly. The stated purpose of releasing the data was to "increase transparency, which helps investors build more accurate credit performance models in support of potential risk-sharing initiatives." (In all, the Fannie and Freddie data used in this study represents some 41 million loans, 1.8 billion quarterly observations, and over $7.2 trillion of origination volume. This is a treasure trove of data is the backbone of my analysis.

HMDA

The GSEs weren't the only mortgage players that were coerced into sharing more data after the housing bubble collapsed. The Home Mortgage Disclosure Act (HMDA) was enacted by Congress in 1975 and was implemented by the Federal Reserve Board's Regulation C. On July 21, 2011, the rule-writing authority of Regulation C was transferred to the Consumer Financial Protection Bureau (CFPB). Regulation C requires lending institutions to report public loan data.

Historically, HMDA data was inaccessible, and was read mostly by policymakers and watchdog groups that were fighting redlining practices. Regulation C requires lending institutions to report sociodemographic data (race, age, gender) for all applicants seeking a loan, as well as approvals and/or denials, and explanations for denials. This data is of particular interest to me because changes in lending behavior has been one of the most prominent hypotheses explaining the housing bubble, and now it has never been reported this thoroughly. Wachter and Levitin (2012) posit that the bubble was a supply-side phenomenon attributable to an excess of mispriced mortgage finance. Mortgage-finance spreads declined and volume increased, even as risk increased—a confluence attributable only to an oversupply of mortgage finance. They then argue that the mortgage-finance supply glut resulted from the failure of markets to price risk correctly due to the complexity, opacity, and heterogeneity of the unregulated private-label mortgage-backed securities (PLS) that began to dominate the market in 2004. The rise of PLS exacerbated informational asymmetries between the financial institutions that intermediate mortgage finance and PLS investors. The new reporting standards only extend back to 2007, and are reported annually (not quarterly). But because the HMDA data captures an aspect of the mortgage market that the GSE data lags behind, it will be included in my analysis.

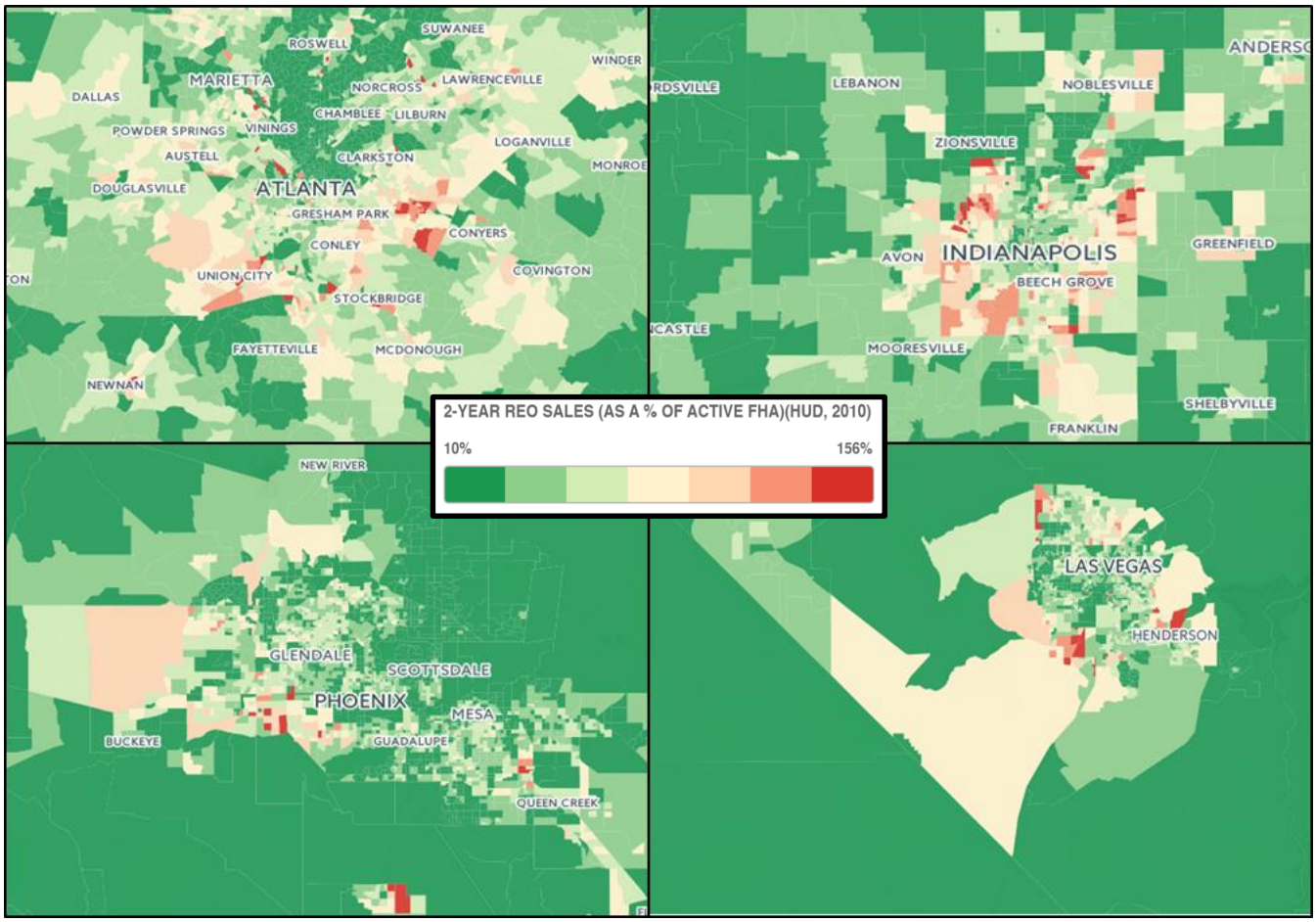
HUD

Under Title III of the Housing and Economic Recovery Act of 2008, HUD's Neighborhood Stabilization Program provided emergency assistance to state and local governments to acquire and redevelop foreclosed properties that might otherwise become sources of abandonment and blight within their communities. The Neighborhood Stabilization Program (NSP) provided grants to every state, certain local communities, and other organizations to purchase foreclosed or abandoned homes and to rehabilitate, resell, or redevelop these homes in order to stabilize neighborhoods and stem the decline of house values of neighboring homes.

There have been three rounds of funding for NSP. The Housing and Economic Recovery Act of 2008 provided a first round of formula funding to States and units of general local government, and is referred to as NSP1. The American Recovery and Reinvestment Act provided a second round of funds in 2009 awarded by competition, and is referred to as NSP2. The third round of funding, NSP3, was provided in 2010 as part of the Dodd-Frank Wall Street Reform Act and was allocated by formula.

While HUD provides robust datasets down to the Census Block Group level from each of its allocation years, this data will not make it into my final analysis, primarily because the data collected by HUD was at the height of the housing bubble and was not reported before, or will be continued. Therefore, it would not be a suitable dataset for running regression and predictive analyses.

HUD NSP3 Foreclosure Rate by Census Block Group, 2010 (Visualized in CartoDB)

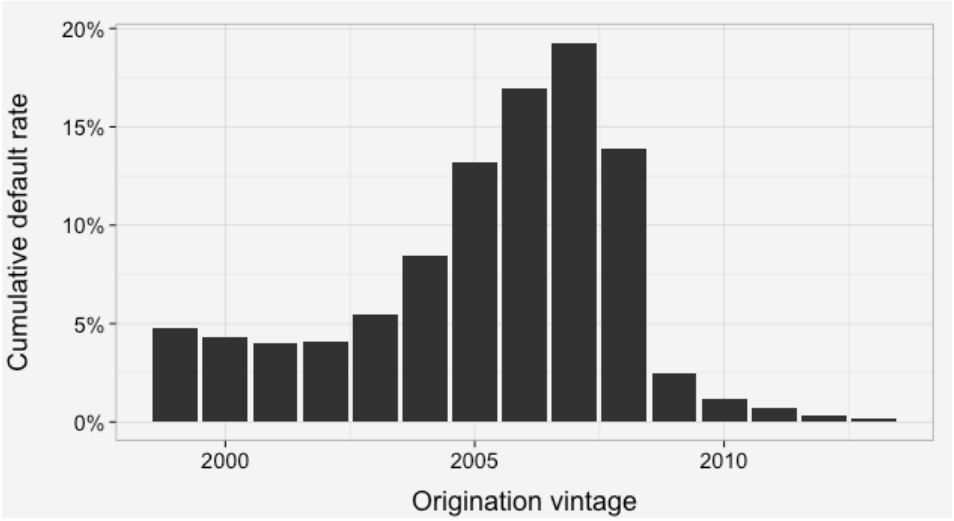


III. Exploratory Analysis

Before diving into entire datasets, I began with some preliminary exploration of the GSE data. Each loan has some static characteristics which never change for the life of the loan - geographic information, the amount of the loan, and a few dozen others. Each loan also has a series of monthly observations, with values that can change from one month to the next, such as the loan's balance, its delinquency status, and whether it was prepaid in full.

I started by calculating simple cumulative default rates for each origination year, defining a "defaulted" loan as one that became at least 60 days delinquent at some point in its life. Note that not all 60+ day delinquent loans actually result in foreclosures, but missing at least 2 payments indicates a "serious delinquency status" by standards defined by HUD.

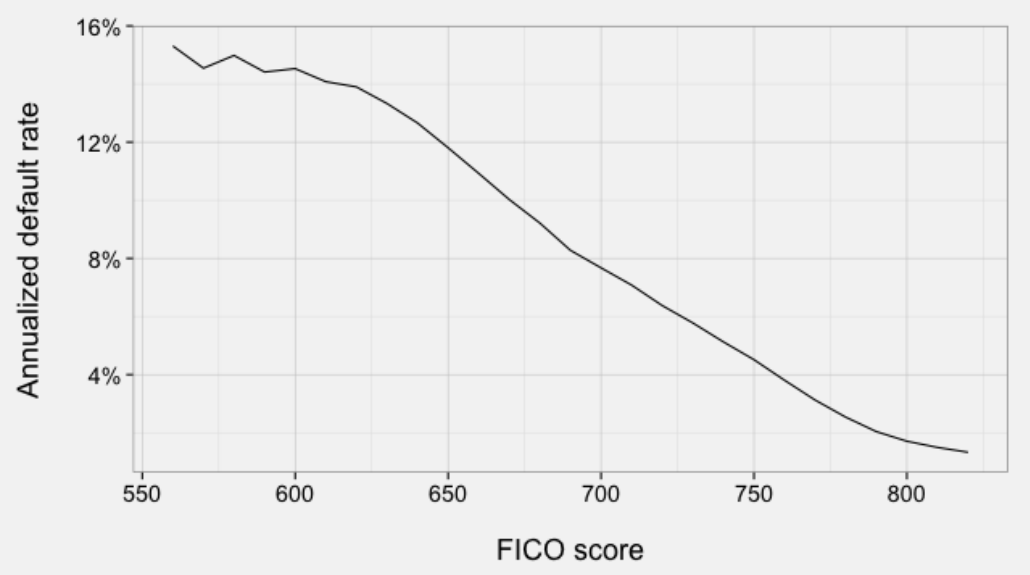
Figure 1: Cumulative Default Rate by Origination Vintage, Fannie Mae & Freddie Mac Loan Performance Data (Generated in R)



About 4% of loans originated from 1999 to 2003 became seriously delinquent at some point in their lives. The 2004 vintage showed some performance deterioration, and then the vintages from 2005 through 2008 show significantly worse performance: more than 15% of all loans originated in those years became distressed. From 2009 through present, the performance has been much better, with fewer than 2% of loans defaulting. Of course, it should be noted that this is at least partially because it takes time for a loan to default, so the most recent vintages will tend to have lower cumulative default rates while their loans are still young.

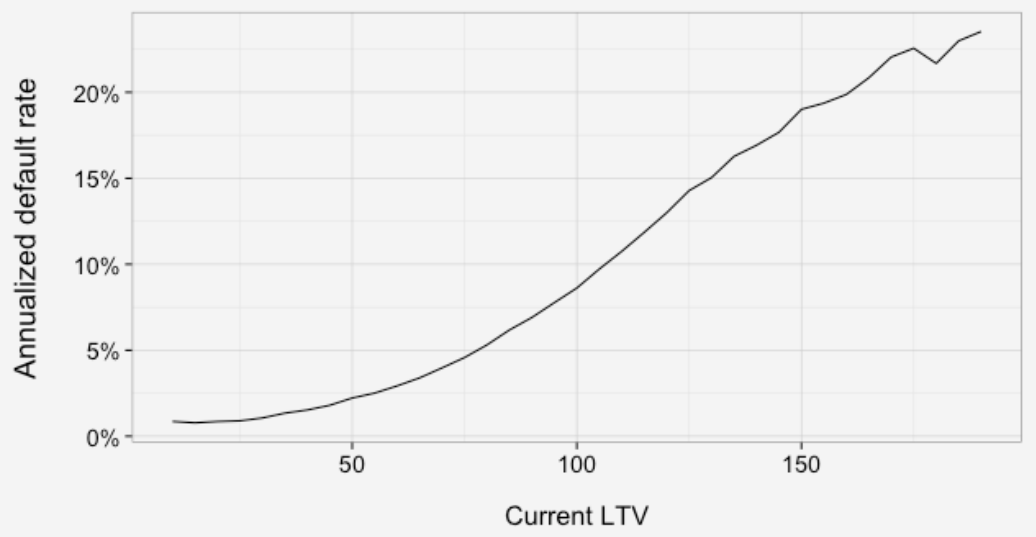
The dataset includes lots of variables for each individual loan beyond geographic location, and many of these variables seem like they should correlate to mortgage performance. Perhaps most obviously, credit scores were developed specifically for the purpose of assessing default risk, so it would be very surprising if credit scores weren't correlated to default rates. Before formulating any specific model, I visualized graphs of aggregated data. I took every monthly observation from 2009-11, bucketed along several dimensions, and calculated default rates. Below is annualized default rate as a function of FICO score.

Figure 2: Annualized Default Rates by FICO Score, Fannie Mae & Freddie Mac Loan Performance Data (Generated in R)



Some of the additional variables include the amount of the loan, the interest rate, the loan-to-value ratio (LTV), debt-to-income ratio (DTI). By simple logic, LTV seemed like another likely predictor of default rates. Because the GSE data only records original LTV figures, I used FHFA's home price data to calculate current loan-to-value ratios for every loan in the dataset. For example, say a loan started at an 80 LTV, but then the home's value has since declined by 25%, while if the balance on the loan has remained unchanged, then the new current LTV would be 0.8 / (1 - 0.25) = 106.7. An LTV over 100 means the borrower is "underwater" – the value of the house is now less than the amount owed on the loan. If the borrower does not believe that home prices will recover for a long time, the borrower might rationally decide to "walk away" from the loan. Not surprisingly, both FICO scores and current LTV ratios are highly correlated to default rate. This is off to a good start, but I believed more explanatory variables were hiding in the data.

Figure 3: Annualized Default Rates by Current LTV, Fannie Mae & Freddie Mac Loan Performance Data (Generated in R)



Don’t You Dare Forget Geography

One cannot possibly discuss housing without hearing the most tired phrase in the real estate industry: “It’s all about location, location, and location.” While that adage is often parroted by pushy realtors trying to sell houses, its holds just as true for the less gleeful outcome of housing transactions. You don’t need to know what personal geodatabase is to know there was a significant spatial component to the outcomes of communities in the economic collapse. The housing downturn was most acute in four states—Arizona, California, Florida, and Nevada—that had experienced some of the highest rates of home price appreciation in the first half of the 2000’s. While these states are not all contiguously located, their similar housing cycles and abundance of either beaches or deserts have led some analysts to label them "Sand States."

There are many well-constructed ideas for why these four states got hit the hardest. For many years, rapid population growth in the Sand States spurred higher than average rates of home construction. Favorable weather and relatively affordable housing are two factors that attracted retirees as well as younger families to these states. In the 1980s and 1990s, population growth rates in Arizona, Florida, and Nevada were between two and four times the national rate. Certain parts of California, such as the Riverside—San Bernardino metropolitan area, experienced similarly high rates of population growth. Rapid population growth continued into the early years of this decade. From 2004 to 2007, Arizona and Nevada ranked as the two fastest growing states in the nation, followed closely by Florida, which ranked ninth.

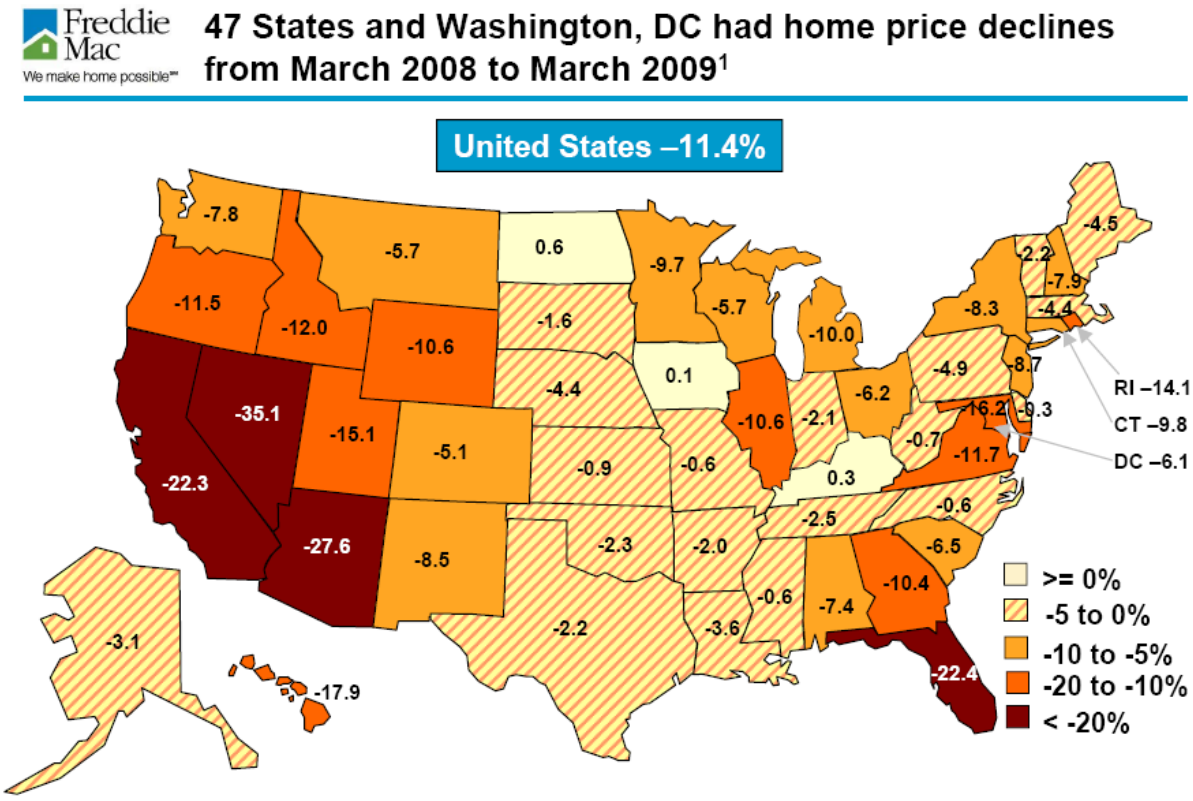


Figure 4: Home Price Declines by State, March 2008-09, Source: Freddie Mac Quarterly Report April 2009

The Sand States' Share of Foreclosure Activity, 2007-2010		
	National Share of Foreclosures Started	National Share of Mortgages Serviced

California	19.20%	12.90%
Florida	16.20%	7.80%
Arizona	4.40%	2.70%
Nevada	2.70%	1.20%
Sand States Total:	**42.50%**	**24.60%**

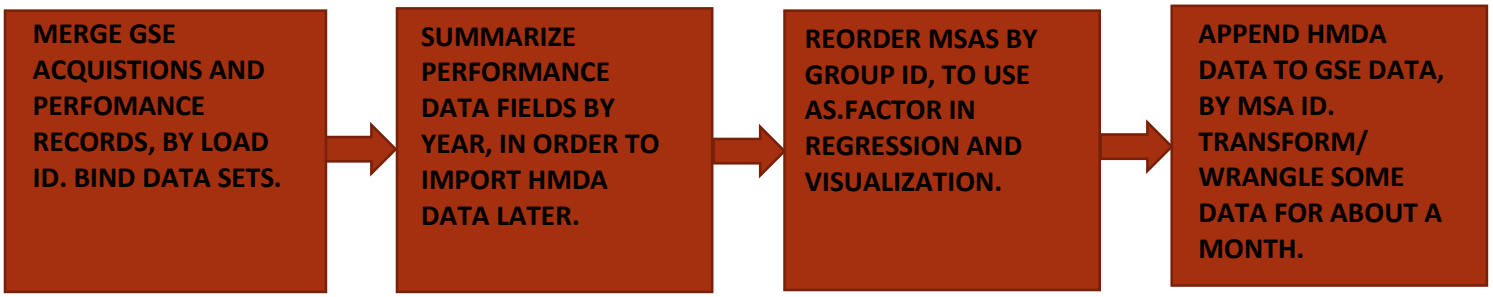
Source: Mortgage Bankers Association

During this decade, strong demand for housing, supported by a growing population and an expanding economy, contributed to growing housing market imbalances across the Sand States. Perhaps the best measure of the imbalances that accumulated in booming housing markets during this decade was the relationship between home prices and incomes. In the years leading up to the housing downturn, escalating home prices far outpaced income growth. A combination of factors drove the housing sector imbalances in the Sand States to unprecedented levels. Under normal market conditions, strained affordability tends to limit housing demand because fewer households can purchase a home using traditional mortgage financing. However, in this cycle, new mortgage "affordability" products were commonly used to finance home purchases. Besides traditional adjustable-rate mortgages (ARMs), affordability products included hybrid ARMs, which have a low, fixed interest rate for several years followed by a market rate that is frequently much higher.

By 2006, nearly half of total U.S. originations of privately securitized affordability mortgages were made in the four Sand States alone. Moreover, the proportion of these mortgages originating in these states, including nontraditional mortgages, rose as home prices escalated. During 2002, these products accounted for roughly half of the privately securitized mortgage originations in each of the Sand States, comparable to the rest of the nation. By 2006, however, the proportion of these products had increased to 80 percent of privately securitized mortgage originations. Nationwide, the percentage was about 70 percent. ARMs are one of the favorite villains for researchers picking apart the pieces of the housing crash. While it’s been proven that ARMs had a significant contribution to the collapse, all the data used in my analysis thus far, and moving forward, is in the much more “safe” and “transparent” primary and secondary markets that deal in 30-year, fixed rate mortgages. If ARMs were the culprit, the GSE’s wouldn’t have hemorrhaged until the point at which they needed divine government intervention.

IV. Formulating a Preliminary Model of Mortgage Default Risk

After performing the preliminary analysis on the GSE data, I feel I have found some robust explanatory variables to begin constructing a regression model with. Firstly, I performed a series of tedious data management steps that I will briefly explain – in graphs.



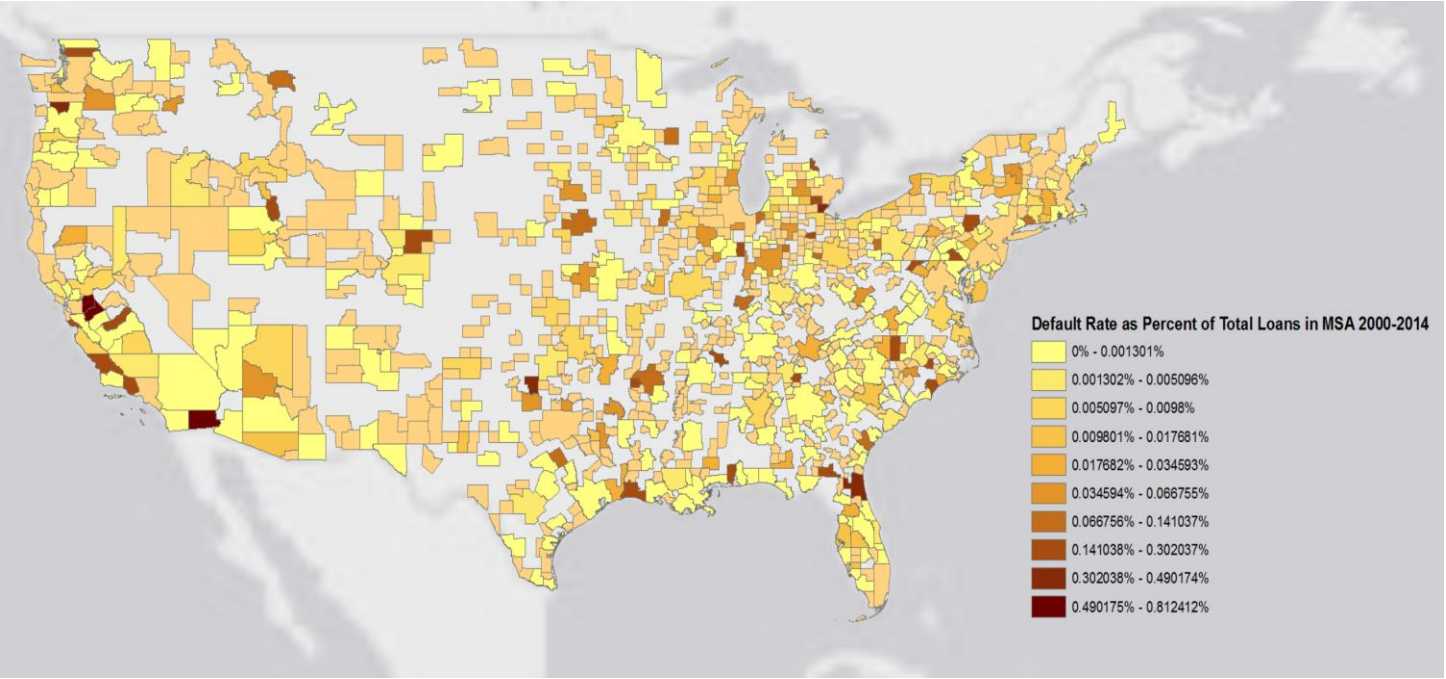
After that brief interlude, I am left with 8 years of data encompassing 324 MSA polygons (urban cores with at least 50,000 people) – totaling 3,024 shapes with data. I am ready to plot a simple linear regression model, taking into account only non-space/time variables. I found the most robust model called mean FICO scores, mean Original Loan Rate, mean Current LTV, mean Loan Approval Ratio, and mean Debt to Income Ratio.[1] Below is the summary output in R environment. I am pleased that all variables are statistically significant, and that the model produces a the high initial R^2 value, even though it is a simple linear run.

```
Call:
lm(formula = Default_Rate ~ meanFICO + meanOriginal_Rate + meanLTV
+ meanLoan_Approval_Ratio   + meanIncome_to_Debt_Ratio, data = MSAData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22727 -0.01304 -0.00229  0.00829  0.32583

Coefficients:
                            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)                 8.975e-01  4.546e-02  19.744  < 2e-16 ***
meanFICO                   -9.809e-04  4.911e-05 -19.972  < 2e-16 ***
meanOriginal_Rate           1.309e-02  8.867e-04  14.761  < 2e-16 ***
meanLTV                    -1.222e-03  1.310e-04  -9.323  < 2e-16 ***
meanLoan_Approval_Ratio    -1.187e-01  1.125e-02 -10.550  < 2e-16 ***
mean_Income_to_Debt_Ratio  -3.708e-02  6.542e-03  -5.667 1.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02976 on 3024 degrees of freedom
Multiple R-squared:  0.5614,   Adjusted R-squared:  0.5605
F-statistic: 493.9 on 5 and 2882 DF,  p-value: < 2.1e-16
```

[1] The final two variables in the model came from the HMDA Data. Loan Approval ratio was calculated by subtracting Loans Approved/Loans Denied from 1. Debt to Income was calculated taking the mean Loan amount divided by the mean Applicant Income.

V. Still Can't Forget About Geography (GWR)

Cumulative Default Rates for Fannie Mae / Freddie Mac Mortgages Originated 2000-2014, by MSA (Generated in ArcMap)



After plotting this map, it is becoming more apparent that there is indeed a significant spatial component to mortgage default patterns. I used the gwr() package in R to perform geographically weighted regression in addition to the original model parameters. GWR is a local form of OLS linear regression that is used to model spatially varying relationships. When objects and their attributes have a high degree of spatial clustering, the GWR model will improve the accuracy of the base model. Not surprisingly, GWR improved the R^2 significantly to 0.796.

```
Call:
gwr(Default_Rate ~ meanFICO + meanOriginal_Rate + meanLTV
+ meanLoan_Approval_Ratio   + meanIncome_to_Debt_Ratio, data = MSAData)
bandwidth = bwG, gweight = gwr.Gauss, hatmatrix = TRUE)

Kernel function: gwr.Gauss
Fixed bandwidth: 1.507437

Summary of GWR coefficient estimates at data points:
                  Min.    1st Qu.     Median    3rd Qu.       Max.  Global
X.Intercept.  2.275e-01  1.032e+00  1.304e+00  1.788e+00  4.324e+00  1.5970
tbl_meanFI   -4.810e-03 -1.954e-03 -1.459e-03 -1.205e-03 -1.800e-04 -0.0017
tbl_meanLT   -7.865e-03 -3.370e-03 -2.390e-03 -1.596e-03 -9.672e-05 -0.0027
tbl_app_rt   -6.914e-01 -1.154e-01 -4.296e-02  1.183e-02  2.282e-01 -0.0877
tbl_appinc   -9.615e-02 -3.562e-03  2.005e-02  5.660e-02  5.539e-01 -0.0254

Number of data points: 3024
Effective number of parameters (residual: 2traceS - traceS'S): 322.3438
Effective degrees of freedom (residual: 2traceS - traceS'S): 2826.656
Sigma (residual: 2traceS - traceS'S): 0.0209554
Effective number of parameters (model: traceS): 237.8715
```

```
Effective degrees of freedom (model: traceS): 2511.128
Sigma (model: traceS): 0.02059993
Sigma (ML): 0.01968851
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): -13269.87
AIC (GWR p. 96, eq. 4.22): -13555.41

Residual sum of squares: 1.065615
Quasi-global R2: 0.79636
```

VI. Temporal Data: Components and Pitfalls

The final component I felt was important to add into a model is a temporal variable. It is impossible to ignore the initial graphs that show mortgage defaults accumulating and decreasing in successive patterns. Much of this is due to the worldwide economic downturn, but I also believe that on economies are more sensitive to mini-booms and mini-busts, and homeowners are more likely to default/walk away from their mortgages if they see their neighbors doing so. Unfortunately, spatiotemporal regression analysis is a very new field, and the statistical packages I explored left me wanting for a something a bit more robust. I will outline my research process. First, just for the sake of experimentation, I plotted a linear model using just the years as explanatory variables for mortgage defaults. While I was not surprised there was a high correlation, I was surprised that the explanatory R^2 was nearly as powerful as my initial model made of categorical variables.

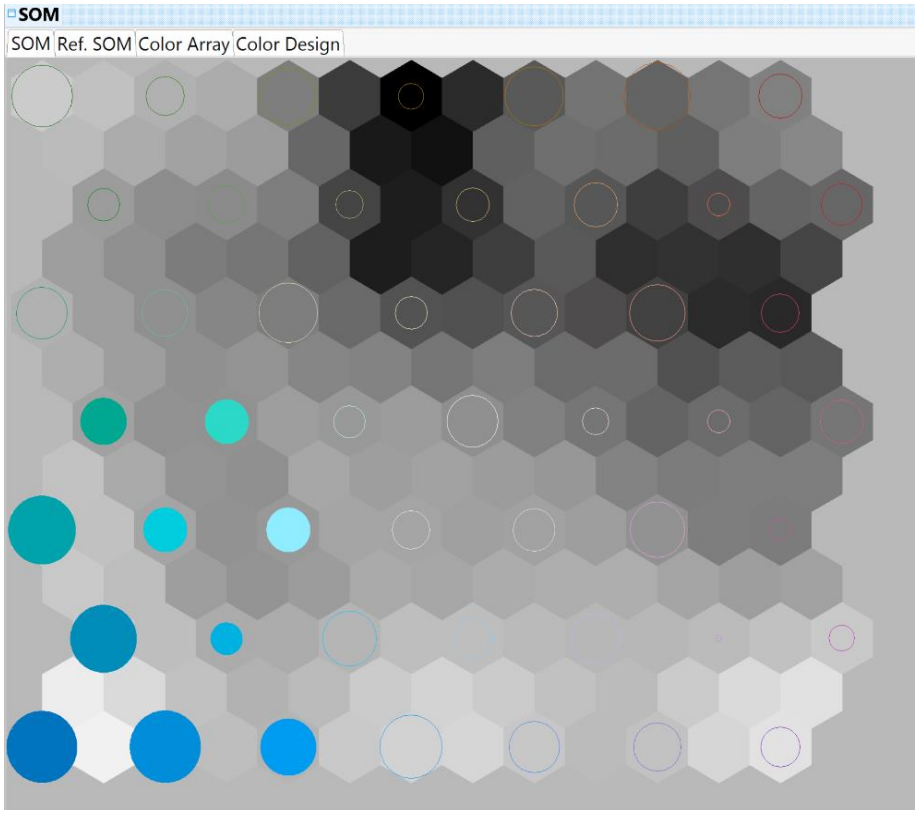
```
Call:
lm(formula = Default_Rate ~ Year, data = MSAData)

Residuals:
      Min        1Q    Median        3Q       Max
-0.088433 -0.003522 -0.000943 -0.000034  0.305448

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.088433   0.001511   58.54   <2e-16 ***
yearfct2008 -0.056397   0.002136  -26.40   <2e-16 ***
yearfct2009 -0.080636   0.002132  -37.82   <2e-16 ***
yearfct2010 -0.084911   0.002130  -39.85   <2e-16 ***
yearfct2011 -0.086665   0.002129  -40.71   <2e-16 ***
yearfct2012 -0.087490   0.002132  -41.04   <2e-16 ***
yearfct2013 -0.088057   0.002152  -40.92   <2e-16 ***
yearfct2014 -0.088399   0.002109  -41.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02858 on 3024 degrees of freedom
Multiple R-squared:  0.5035,   Adjusted R-squared:  0.5023
F-statistic: 417.2 on 7 and 2880 DF,  p-value: < 2.2e-16
```

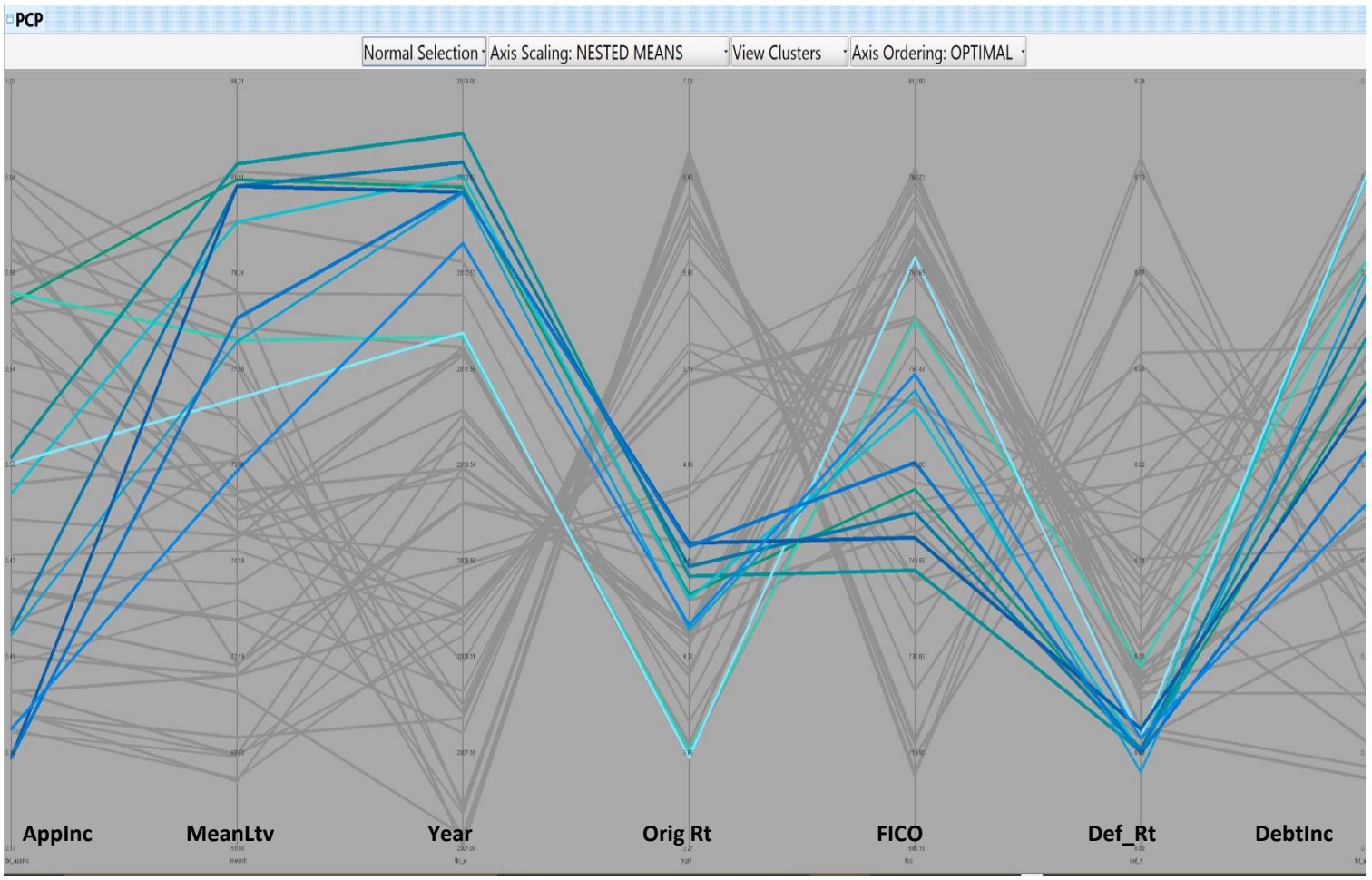
Visualizing Spatiotemporal Data

While I had no luck finding the ideal package or technique to analyze the temporal aspect of my data, I discovered VIS-STAMP software as a mean of visualizing it in an analytical setting. The software tool integrates computational, visual, and cartographic methods together to detect and visualize multivariate spatio-temporal patterns. It is able to: perform multivariate clustering and abstraction (including time-series clustering) with a Self Organizing Map (SOM); encode SOM result with colors derived from a two-dimensional diverging-diverging color scheme; visualize the multivariate patterns with an enhanced Parallel Coordinate Plot (PCP) display, which serves as a multivariate "legend" in the integrated system; visualize the spatio-temporal variations of multivariate patterns, or the space-variable variations of temporal patterns in a hierarchical, computationally sortable matrix and a temporally or geographically ordered map matrix. VIS-STAMP creates the SOM based on values in detects cluster together based on space, time, and value. The user can select hexagonal quadrants form the SOM to see the data display in the other visualizations. The visualizations produced by VIS-STAMP follow.

In the first three plots, the SOM, PCP and Map Matrix are clusters selected from the SOM which have low default rate matters. PCP variables self-organized on the X-axis based on correlation. The Y-axis shows variable values, low to high. Which makes it not surprising in the PCP that Applicant Income and FICO scores cluster high, while Loan Rates cluster low. It should also be noted that the Map Matrix layer didn't even have the spatial weight to draw shapefiles for the years 2007-2009, which is kind of temporal proof I was searching for.

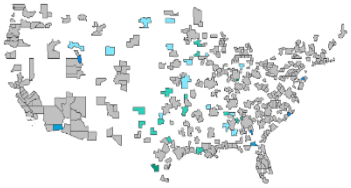
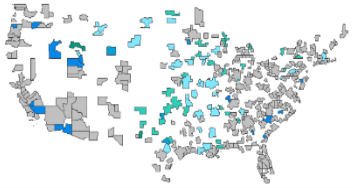
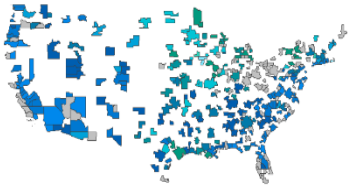
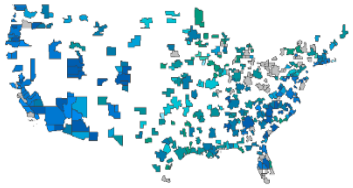
The next three images are in the same sequence, but they are clusters of high default rates. In contrast, high rates of defaults have the highest clustering where FICO scores are low, and in the years 2007-2009.

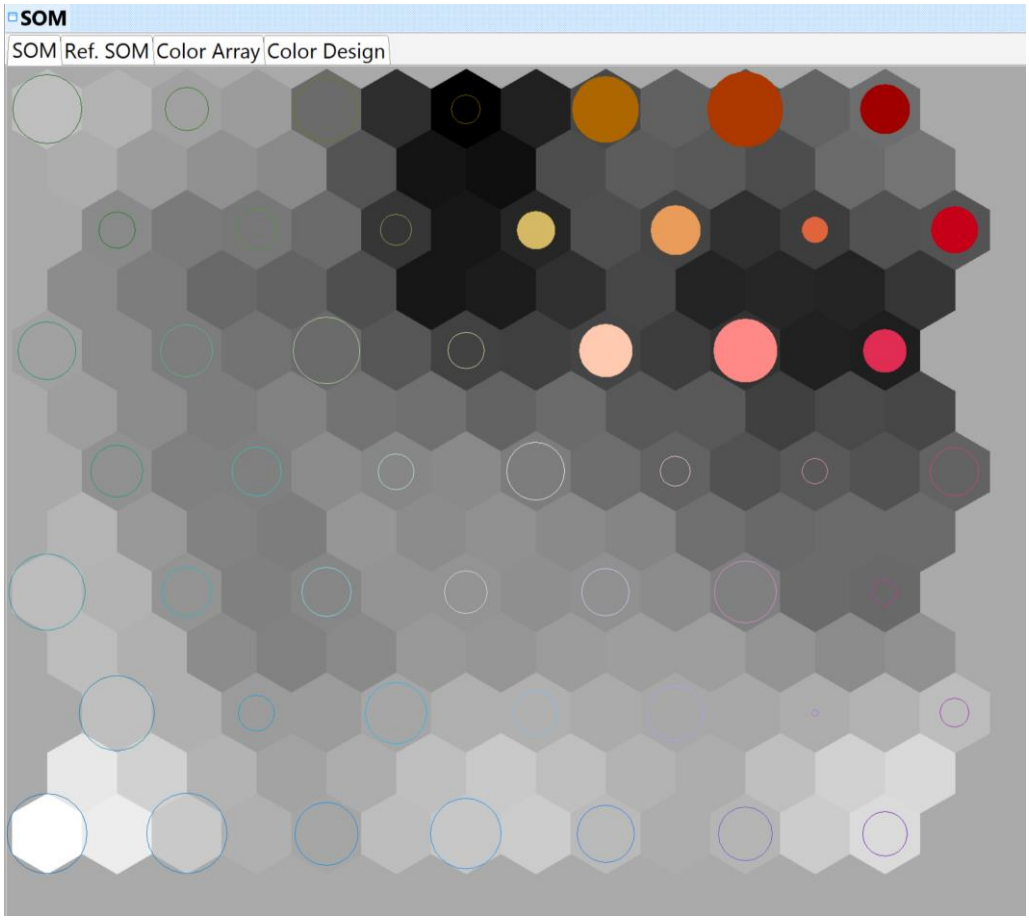

SOM
SOM
Ref. SOM
Color Array
Color Design

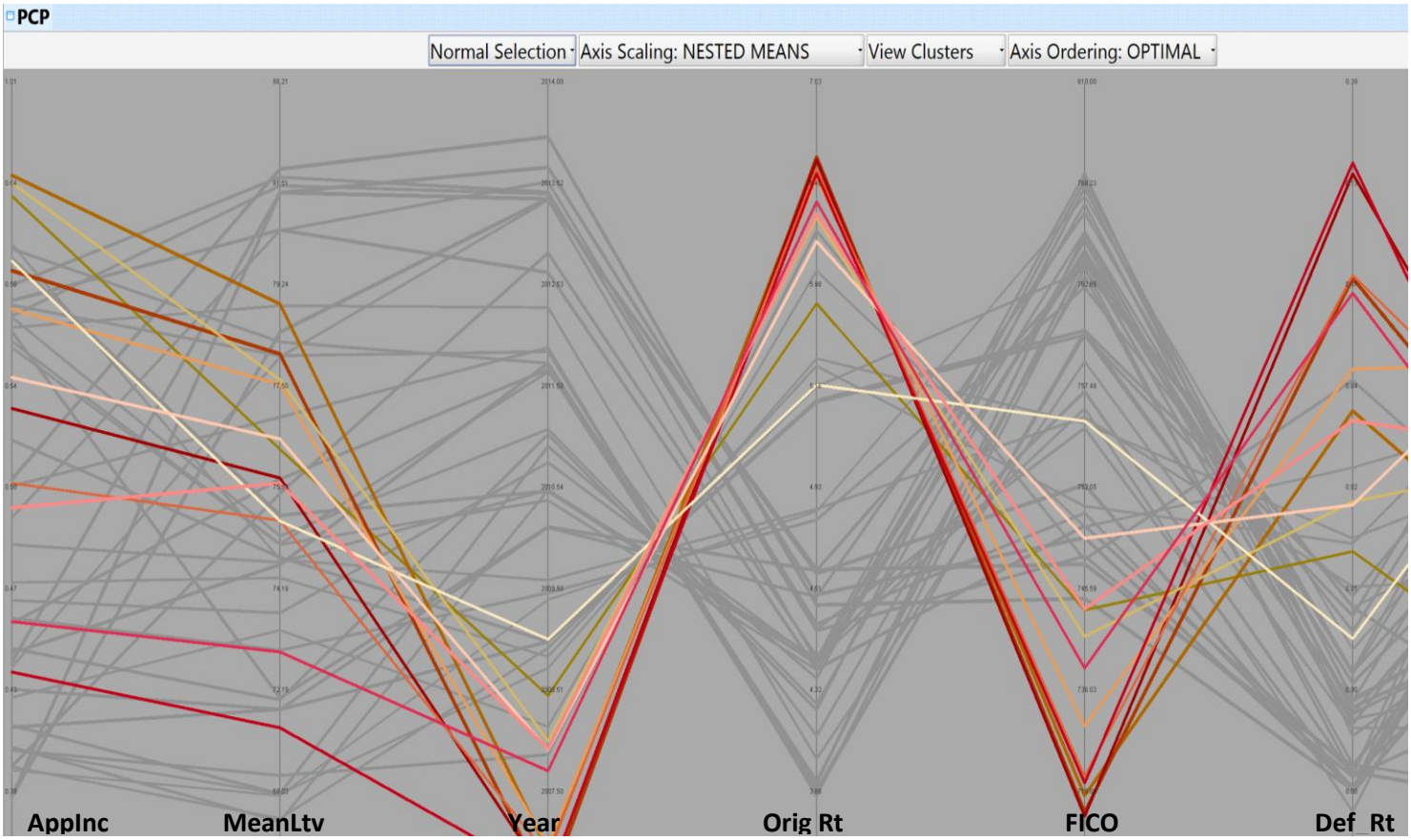

PCP
Normal Selection
Axis Scaling: NESTED MEANS
View Clusters
Axis Ordering: OPTIMAL
AppInc
MeanLtv
Year
Orig Rt
FICO
Def_Rt
DebtInc

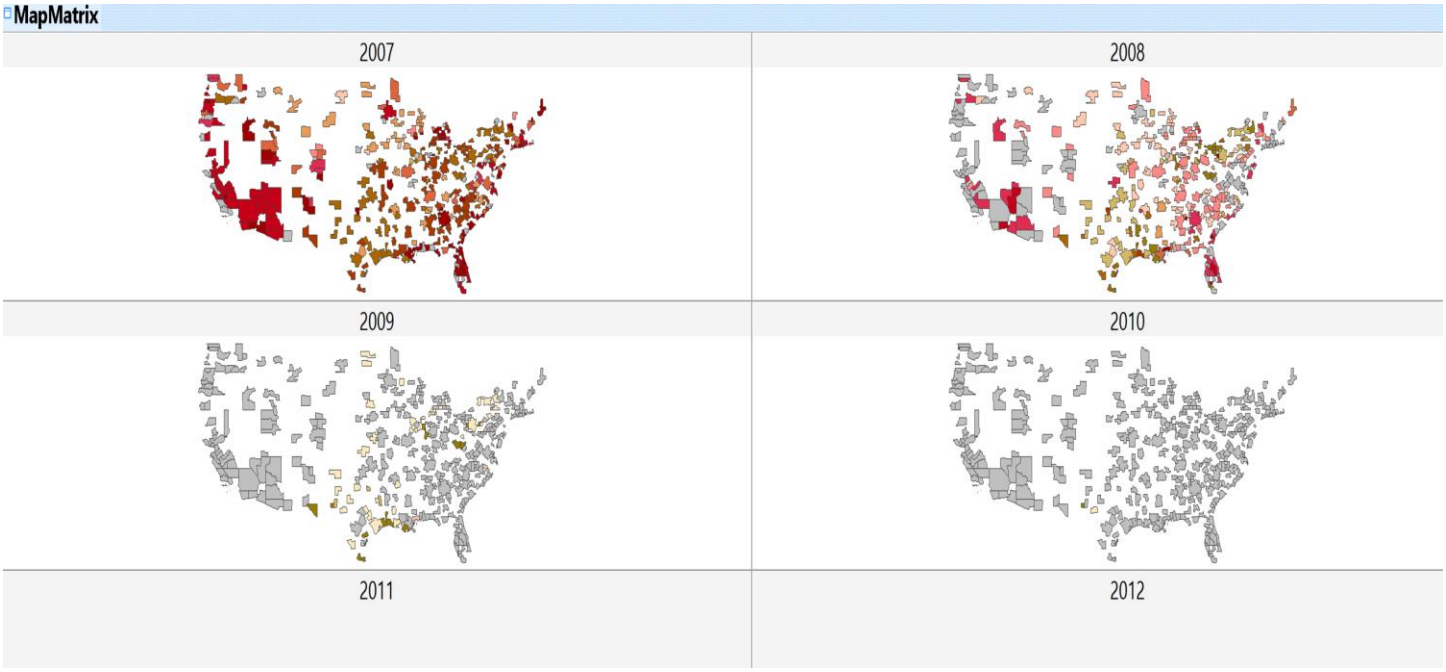
MapMatrix

2007	2008
2009	2010
2011	2012
2013	2014


SOM
SOM
Ref. SOM
Color Array
Color Design


PCP
Normal Selection
Axis Scaling: NESTED MEANS
View Clusters
Axis Ordering: OPTIMAL
AppInc
MeanLtv
Year
Orig Rt
FICO
Def_Rt



Settling on a Temporal Model

After much digging and trialing packages in R, I settled on a Cox proportional hazards model. This model helps give a sense of which variables have the largest relative impact on default rates, while also being capable of handling a time variable. The model assumes that there's a baseline default rate (the "hazard rate"), and that the independent variables have a multiplicative effect on that baseline rate. Most commonly, this examination entails the specification of a linear-like model for the log hazard. For example, a parametric model based on the exponential distribution may be written as

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

or, equivalently,

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

that is, as a linear model for the log-hazard or as a multiplicative model for the hazard. Here, i is a subscript for observation, and the xs are the covariates. The constant α in this model represents a kind of log-baseline hazard, since log hi(t) = α [or hi(t) = e α] when all of the xs are 0. There are similar parametric regression models based on the other survival distributions described in the.

The Cox model, in contrast, leaves the baseline hazard function α(t) = log h0(t) unspecified:

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

```
coxph(formula = formula, data = all)

  n= 3082, number of events= 3024

               coef exp(coef)  se(coef)      z Pr(>|z|)
tbl_meanFI  0.029364  1.029799  0.001788 16.427  < 2e-16 ***
tbl_meanLT -0.009675  0.990372  0.006317 -1.532    0.126
```

```
tbl_app_rt  2.623027 13.777359  0.440422  5.956 2.59e-09 ***
tbl_appinc  1.927662  6.873423  0.257295  7.492 6.78e-14 ***
tbl_yr      0.504635  1.656381  0.015221 33.154  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
tbl_meanFI    1.0298    0.97106    1.0262     1.033
tbl_meanLT    0.9904    1.00972    0.9782     1.003
tbl_app_rt   13.7774    0.07258    5.8114    32.663
tbl_appinc    6.8734    0.14549    4.1511    11.381
tbl_yr        1.6564    0.60373    1.6077     1.707

Concordance= 0.862  (se = 0.011 )
Rsquare= 0.628   (max possible= 1 )
Likelihood ratio test= 2719  on 5 df,   p=0
Wald test            = 2126  on 5 df,   p=0
Score (logrank) test = 2442  on 5 df,   p=0
```

VII. Results

While powerful, the Cox hazards model could not match the explanatory analysis of the GWR. While I am satisfied the accepting the GWR, I look forward to investigating and better understanding spatiotemporal regression analyses, in hopes of developing working models or submodels that can process regression through space and time. In working with and being befuddled by the temporal data, I realized that the housing crash was an exceptional event, and the default rates from that time were (hopefully once in a lifetime). Likewise, the housing market is still in a recovery phase from the crash, and defaults today are exceptionally low. Thus even if there was a properly working spatiotemporal model, data from the past decade would likely confound it.

For now, GWR is a powerful tool for regressing mortgage default risk and its covariables. If the trend of data disclosure grows, data sets with higher resolutions (neighborhood, block group level) could help develop more precise GWR models. While the record number was enormous, the GSE data dump only represents fifteen years of mortgage performance observations. As time progresses and more federal and private enterprises are pushed to release more analytical data, the research possibilities become endless.

VIII. Conclusions

As the eight-year anniversary of the collapse of Lehman Brothers nears, there is still little consensus as to what cause the housing bubble and ensuing Great Recession that the global economy is just now starting to pull out of. The collapse in housing prices and construction constituted one of the most dramatic such episodes in the history of the U.S. housing industry. The episode entailed a large decrease in credit standards, increased use of subprime lending, billions in lost home equity, and a consequent decline consumption spending. In addition, the post-collapse financial distress decimated lender balance sheets, and the high level of foreclosures appear to have a negative effect even on the value of neighboring homes whose mortgages were not in risk of default (Rogers and Winters, 2009).

Accordingly, many researchers have attempted to sort out the causes of the run up and crash in housing. There have been numerous factors proposed as being the cause of the crisis, from lax regulation to tax code changes to irrational consumer expectations of rising housing prices. None of these explanations, however, is capable of fully explaining the housing bubble. Neither is this project.

In reality, the housing bubble and subsequent financial crisis were the result of a confluence of macro- and micro-economic phenomena, upon which academics have assigned varying responsibility. This does not mean the crisis can pass without gleaning lessons which will hopefully prevent a future housing crash of such magnitude. One of the silver linings of the economic collapse and subsequent maneuvers made by federal regulatory and housing-related agencies is the open dissemination of an unprecedented amount of data related to housing markets, lending behavior, and default patterns. One of the most alarming aspects of the housing collapse was that barely anyone saw it coming, and those who did either weren't listened to (or decided to quietly profit off it). What I hope this project accomplishes is to make a contribution to the data dissemination – to put primary source data into visualizations and models that can be replicated and improved by others interested protecting the health of the U.S. housing market.

As mentioned in the Introduction, while the scope of this project was large, the ambitions were humble. There has been a staggering amount of housing- and mortgage-related data that has been made available for public consumption in the past several years, researchers have only just begun sifting through the massive data sets and formulating conclusive studies. While I had come upon several studies in my research that had aggregated the mass data sets from the GSEs into analyses, resources that took into account geography were scant, and ones that combined GSE data with other federal data sets were nonexistent. Geographic regression models have become popular in interpolating house prices, and I believe that similar models can be used to interpolate default risk as well. There will always be a human-level decision to be made when it comes to buying a house or defaulting on a mortgage balance, but as with everything in real estate – location matters.

Data Sources

Fannie Mae Single Family Loan Performance Data. http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html

Federal Housing Finance Agency House Price Index. http://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx

Freddie Mac Single Family Loan Data Set. http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

HMDA Data Tables. http://www.consumerfinance.gov/data-research/hmda/

HUD Neighborhood Stabilization Program Data. https://www.huduser.gov/portal/datasets/NSP.html

References

Belsky, Can. “A Primer on Geographic Information Systems in Mortgage Finance” 1998. Journal of Housing Research 9 (1). http://content.knowledgeplex.org/kp2/img/cache/documents/1164.pdf

Chatterjee, Satyajit and Burcu Eyigungor. "A Quantitative Analysis of the US Housing and Mortgage Markets and the Foreclosure Crisis." Federal Reserve Bank of Philadelphia. March 2015. https://philadelphiafed.org/research-and-data/publications/working-papers/2015/wp15-13.pdf

Cohen, Ioannides. “Spatial Effects and House Price Dynamics in the USA” (2016). Journal of Housing Economics 31. https://sites.tufts.edu/yioannides/files/2015/02/CohenIoannidesThanapisitikul-REVISED-2-17-15.-Fullpdf.pdf

Hepp, Salma. “Foreclosures and Metropolitan Spatial Structure: Establishing the Connection” (2013). Housing Policy Debate 23 (3). http://drum.lib.umd.edu/bitstream/handle/1903/11976/Hepp_umd_0117E_12539.pdf?sequence=1&isAllowed=y

Li, Yanmei. “Geography of Opportunity and Residential Mortgage Foreclosure: A Spatial Analysis of a U.S. Housing Market” (2011).Journal of Urban and Research Analysis (3) 2. http://www.jurareview.ro/2011_3_2/a_2011_3_2_5_li.pdf

Ogbe, Adejoh Emmanuel, "Spacial analysis of foreclosure and neighborhood characteristics in Miami metropolitan area, Florida" (2015). Electronic Theses and Dissertations. Paper 178. http://scholarworks.uni.edu/etd/178

Rogers, William and William Winter. “The Impact of Foreclosures on Neighboring Homes” (2009). Journal of Real Estate Research. http://pages.jh.edu/jrer/papers/pdf/past/vol31n04/04.455_480.pdf

Rugh, Albright & Massey. “Race, Space, and Cumulative Disadvantage: A Case Study of the Subprime Lending Collapse” (2015). Oxford Journal of Social Problems. http://drum.lib.umd.edu/bitstream/handle/1903/11976/Hepp_umd_0117E_12539.pdf?sequence=1&isAllowed=y

Wachter, Susan and Adam Levitin. “Explaining the Housing Bubble” (2012). Georgetown Law Journal, Vol. 100, No. 4. http://georgetownlawjournal.org/files/2012/04/LevitinWachter.pdf

Zhang, Duan. “Spatial Dependence and Neighboorhood Effects in Mortgage Lending: A Geographically Weighted Regression Approach” (2006). Luck Center for Real Estate. http://lusk.usc.edu/sites/default/files/working_papers/wp_2006-1007.pdf