Mortgage data description

The dataset contains conforming mortgage loans purchased by Fannie Mae and Freddie Mac over 2000–2014. To focus on a homogeneous product, we restrict our sample to single family, 30-year fixed rate, purchase mortgages for primary residency purposes.

Variables

Variable	Description	Value
source	The government entity that purchases the loan.	FN: Fannie Mae; FD: Freddie Mac
loan_id	Unique identifier assigned to each loan.	12-digit string
Year_orig	The year of note origination.	numeric, no missing values
Quarter_orig	The quarter of note origination.	numeric, no missing values
delinquent30	Dummy variable which is 1 if the loan has ever been at least 30 days past due from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
frst_dte	The date on which the first full month of interest begins to accrue.	numeric, no missing values, MM/01/YYYY
orig_rt	The original note rate as indicated on the martgage note.	numeric, no missing values
orig_amt	The unpaid balance of the mortgage on the note origination date, rounded to the nearest $1,000.	numeric, no missing values
orig_trm	The number of scheduled monthly payments of the mortgage. In this sample, we only include 30-year mortgages.	360
oltv	Original loan-to-value, dividing the original mortgage loan amount on the note date by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	numeric, in percentage points, no missing values.

ocltv	Original combined loan-to-value, dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount by the lesser of the mortgaged property's appraised value on the note date or its purchase price.	numeric, in percentage points, no missing values.
dti	Original debt-to-income ratio, dividing the sum of the borrower's monthly debt payments by the total monthly income used to underwrite the loan as of the date of the origination.	numeric, in percentage points, no missing values
cscore_b	Credit score at the origination date.	numeric, no missing values
mi_pct	Mortgage insurance coverage, the percentage of loss coverage on the loan in case of default provided by a mortgage insurer. Usually non-zero for loans with LTV greater than 80.	numeric, in percentage points, no missing values
fthb_flg	First time homebuyer flag, an indicator for whether the borrower is an individual who had no ownership interest in a residential property during the three-year period preceding the date of the purchase of the mortgaged property.	Y: Yes; N: No
num_bo	Number of borrowers, categorical variable, the number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property.	1 if there is 1 borrower; 2 if there are more than 1 borrowers
purpose	Indicates whether the mortgage loan is a refinance mortgage or a purchase mortgage. In this sample, we only include purchase mortgages.	P: purchase mortgage

prop_typ	Property type, denotes whether the property is a condominium, leasehold, planned unit development, cooperative share, manufactured home, or single family home. In this sample, we only include single family home.	SF: 1–4 fee simple
num_unit	Number of units, denotes whether the mortgage is a one-, two-, three-, or four-unit property.	numeric, 1 to 4, no missing values
occ_stat	Occupancy status, denotes whether the mortgage type is owner occupied, second home, or investment property. In this sample, we only include owner occupied property.	P: primary residence
state	A two-letter abbreviation indicating the state within which the property securing the mortgage is located.	2-digit string, no missing values
zip_3	The postal code for the location of the mortgaged property, only the first three digits of the 5-digit postal code are disclosed.	3-digit string, no missing values
cd_msa	This is based on the designation of the Metropolitan Statistical Area or Metropolitan Division based on 2010 and 2013 census. Missing value indicates that the area in which the mortgaged property is located not an MSA/MD.	5-digit integer, possible missing values

Variables below may not be useful but included for potential interest.

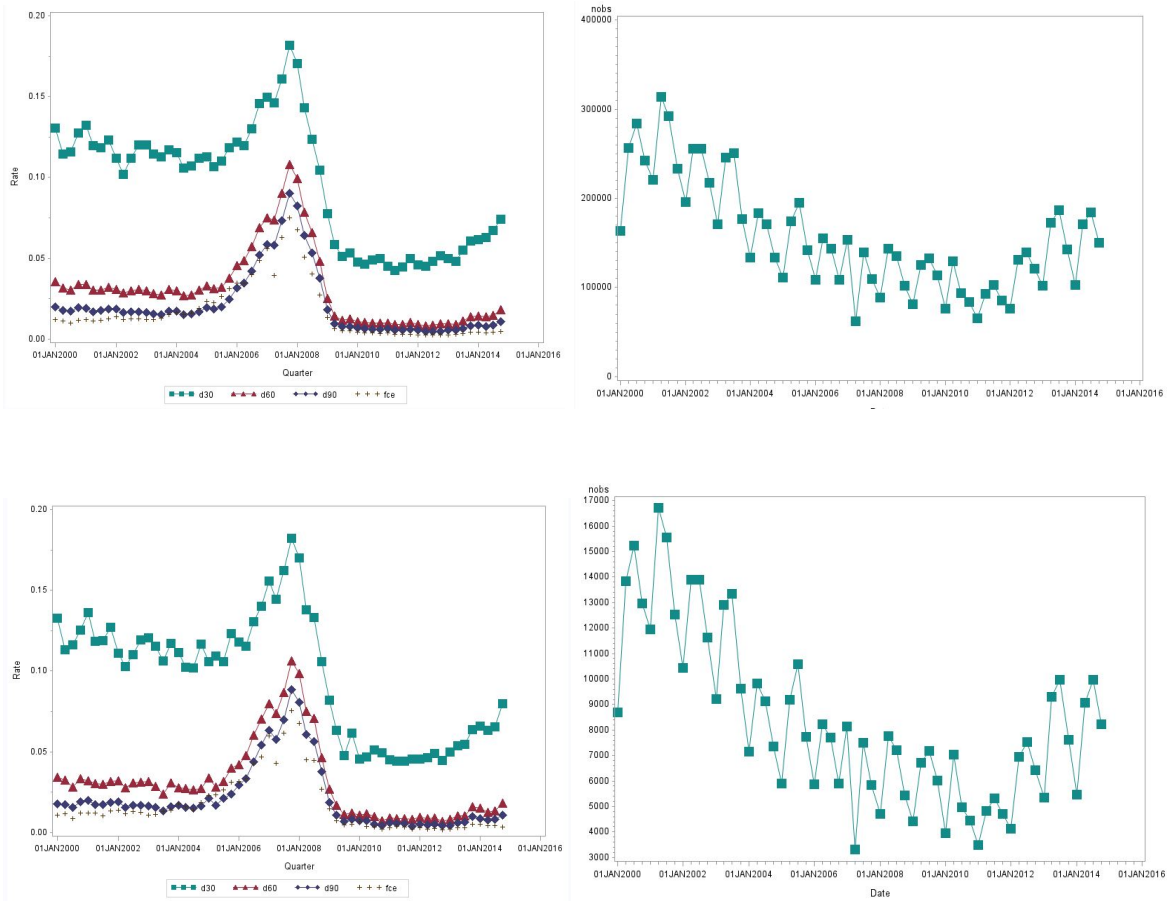
Variable	Description	Value
orig_chn	Origination channel, indicates whether a Broker or Correspondent originated or was involved in the origination of the mortgage loan. Missing variable indicates that the channel is unknown.	R: retail; B: broker; C: correspondent; T: third party origination not specified

seller	The lender who sells the loan to GSE, only identified for the lenders who originate more than one percent of total volume within a given quarter).	string
delinquenct60	Dummy variable which is 1 if the loan has ever been at least 60 days past due from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
delinquenct90	Dummy variable which is 1 if the loan has ever been at least 90 days past due from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
foreclosure	Dummy variable which is 1 if a foreclosure happened from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
prepaid_cnt	Dummy variable which is 1 if the loan is prepaid during the period from the date on which the first full month of interest begins to accrue till 3 years after, and 0 otherwise.	numeric, no missing values
f30_dte	The last observed date of at least 30 days past due during the period from the date on which the first full month of interest begins to accrue till 3 years after.	numeric, MM/01/YYYY
f60_dte	The last observed date of at least 60 days past due during the period from the date on which the first full month of interest begins to accrue till 3 years after.	numeric, MM/01/YYYY

f90_dte	The last observed date of at least 90 days past due during the period from the date on which the first full month of interest begins to accrue till 3 years after.	numeric, MM/01/YYYY
fce_dte	The last observed date of foreclosure during the period from the date on which the first full month of interest begins to accrue till 3 years after.	numeric, MM/01/YYYY

Delinquency rate

The full sample contains 9,330,173 loans in 2000–2014 period. The quarterly delinquency rate and number of loan originations are plotted in the top two graphs. For our exercise, we draw a random sample of 500,000 loans. And the quarterly delinquency rate and number of loan originations of this subsample are plotted in the bottom two graphs.



Macro data

Data file	Description	Unit of observation
hpi_state.csv	The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties.	state-by-quarter level
hpi_msa.csv	The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties.	msa-by-quarter level
income_state.csv	The Quarterly Census of Employment and Wages (QCEW) program publishes a quarterly count of employment and wages reported by employers covering more than 95 percent of U.S. jobs. Weekly income indicates the average weekly income in a given quarter and geographic area.	state-by-quarter level
income_msa.csv	The Quarterly Census of Employment and Wages (QCEW) program publishes a quarterly count of employment and wages reported by employers covering more than 95 percent of U.S. jobs. Weekly income indicates the average weekly income in a given quarter and geographic area.	msa-by-quarter level

unemployment_state.csv	The Local Area Unemployment Statistics (LAUS) program is a federal-state cooperative effort in which monthly estimates of total employment and unemployment are prepared.	state-by-month level
unemployment_msa.csv	The Local Area Unemployment Statistics (LAUS) program is a federal-state cooperative effort in which monthly estimates of total employment and unemployment are prepared.	msa-by-month level
rate.csv	The dataset downloaded from St. Louis Fed FRED contains 30-year fixed rate mortgage rates and 3 month treasury bill rates.	monthly level

Notes:

- In the mortgage dataset, there are cases where the mortgaged property is not located in an MSA/MD. In these case, you might consider to use state-level data as a proxy.
- In the mortgage dataset, we observe the quarter of origination and the month of first scheduled payment, but not the month of the origination. On average, there is a 45-day or two-month gap between mortgage closing and the first payment. When using macro variables, you might want to be careful about what information is already known at the time of prediction.