BANK OF ENGLAND

Staff Working Paper No. 703

A tiger by the tail: estimating the UK mortgage market vulnerabilities from loan-level data

Chiranjit Chakraborty, Mariana Gimpelewicz and Arzu Uluc

December 2017

BANK OF ENGLAND

Staff Working Paper No. 703

A tiger by the tail: estimating the UK mortgage market vulnerabilities from loan-level data

Chiranjit Chakraborty,[1] Mariana Gimpelewicz[2] and Arzu Uluc[3]

Abstract

Following the global financial crisis, macroprudential regulators in a number of countries took actions to mitigate risks arising from stressed mortgage markets to financial and economic stability. Having disaggregated information on the stock of mortgages allows policymakers to analyse particular cohorts of the market that may be more vulnerable to stress, and model how these cohorts may evolve in the future and might affect the outlook for financial and economic stability. To this end, we produce the first ever estimate of the current stock of all regulated UK mortgages at the level of individual loans using data from the flow of new mortgages. We use loan-level information of 14 million UK mortgages at the point each loan was originated or re-mortgaged. Using a series of algorithms from Computer Science, we identify individual loans in the flow of lending that are likely to be still in the stock at different points in time. Then we estimate how key characteristics of mortgages (including borrower incomes, house prices and outstanding loan amounts) are likely to have evolved over time since origination. We validate our overall model by comparing key variables to information available from other sources that provide partial characteristics of the stock, including household surveys and regulatory returns. Our stock estimate suggests that there may have been more vulnerable borrowers in recent years than household surveys suggest. Finally, we illustrate the type of cohort analysis that can be done using the loan-level estimate.

Key words: Mortgage market, housing market, matching, loan-level data, stock model.

JEL classification: D04, E24, G21, R20, R21, R23, R31.

(1) Bank of England. Email: chiranjit.chakraborty@bankofengland.co.uk
(2) Bank of England. Email: mariana.gimpelewicz@bankofengland.co.uk
(3) Bank of England. Email: arzu.uluc@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. The authors would like to thank Andy Haldane, Nicola Anderson, Sujit Kapadia, Paul Robinson, Vasileios Madouros, Gavin Wallis, David Bholat, Rhiannon Sowerbutts, Mette Nielsen, Jochen Schenz, Robert Sturrock, Matthew Thompson, Miao Kang, Philippe Bracke, Philip Bunn, Sagar Shah, and participants of a seminar at the Bank of England for their valuable comments and support. Any errors and omissions, of course, remain the fault of the authors.

The Bank's working paper series can be found at www.bankofengland.co.uk/news/publications

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

Contents

1 Introduction **2**

2 Estimation Approach **4**
2.1 Identifying and removing loans that are not likely to be in the stock . . 5
2.2 Projecting variables of interest forward 9

3 Validation **11**

4 Results **15**
4.1 Loan-level stock distributions 15
4.2 Analysis on home-movers and re-mortgagors 19

5 Conclusion **22**

Bibliography **23**

6 Appendix **25**
6.1 FCA flow data preparation 25
6.1.1 Removing duplicates and aggregating multiple loan entries . . . 25
6.1.2 Imputing missing values 25
6.1.3 Descriptive Statistics 26
6.2 Performance matrix to identify and remove loans that are most likely to be defaulted 27
6.3 Identifying previous mortgages of home-movers that have been paid-off to buy a new property 27
6.4 Identifying loans which have been re-mortgaged 29
6.5 Borrower income projection 29
6.6 Outstanding loan amount projection 31
6.7 Property Value Projection 33
6.8 Validation: FCA stock data and Household Surveys 34
6.8.1 FCA stock data 34
6.8.2 Household Surveys 35

1 Introduction

The global financial crisis in 2007-09 highlighted the implications a stressed mortgage market can have on financial and economic stability. Since then, macroprudential regulators have turned their attention to the housing market and a number of countries have taken actions to mitigate risks to lenders and the wider economy arising from this sector (Bank of England (2014)). Risks to lenders can arise from losses on mortgage lending, which often makes up a significant proportion of their balance sheets (Bank of England (2013)). This can threaten solvency if the lenders are not adequately capitalised to mitigate these losses. Moreover, highly indebted households are more vulnerable to adverse shocks to income and interest rates. Recent research shows that more leveraged mortgagors cut their consumption more sharply in the face of a shock (Mian et al. (2013); Bunn and Rostom (2015)). This can lead to sharp falls in aggregate spending which, in turn, can have an impact on wider economic stability.

Policy makers rely on data to make decisions and calibrate policy to mitigate these risks. Some of the risks arising from the UK housing market to financial and economic stability are best addressed by looking at particular segments of the mortgage market. Aggregated data on the stock only provide a partial picture, limiting the possibility of exploring vulnerable cohorts of borrowers such as those with high loan-to-value (LTV) or loan-to-income (LTI) loans or debt servicing costs. Household surveys can provide some insights into the characteristics of existing mortgage borrowers, but they are often based on small samples and can be subject to reporting bias.

Having a disaggregated estimate of the stock of mortgages allows us to analyse particular cohorts of the market that may be more vulnerable to stress conditions, and also model how they may evolve in the future and might affect the outlook for financial and economic stability. The Financial Conduct Authority (FCA) has been collecting loan-level data on the flow of new regulated lending since 2005 (known as Product Sales Data 001, referred to as 'flow data' hereafter). These data provide insight into characteristics of new mortgages and are useful for modelling how policies are likely to affect new borrowers. For example, using flow data the FPC calibrated its policy to limit the share of new loans with LTIs greater than 4.5 to guard against risks from a growing tail of highgly indebted borrowers. However, we need data on the stock of existing loans in order to understand how policies impact the mortgage market as a whole, as ultimately it is the stock of vulnerable borrowers that could potential pose a risk to financial stability - not just the flow.

The FCA's disaggregated data on loan performance for the stock of outstanding UK mortgages only became available in 2015 (known as Product Sales Data 007, referred to as the 'stock data' hereafter). These stock data are a helpful step forward but have major limitations. They do not include information on borrowers' current incomes or property values which are necessary variables to calculate key metrics such as LTV,

LTI and debt servicing ratios (DSR). While waiting for the FCA's stock data to build up a time series from 2015 onwards, we can exploit the existing flow data to estimate the historical stock, which are richer in characteristics and include a boom-bust cycle. To this end, we develop one of the first approaches to estimate the stock of mortgages at the loan-level from the flow data. This involves tracking the life-cycles of individual mortgages and projecting forward variables of interest.

The contributions of our work are twofold. First, the use of loan-level data to model housing market dynamics, and understand or simulate the impact of macroprudential tools remain scarce (Aron and Muellbauer (2010)). Other central banks have recently increased their efforts.[1] However, to the best of our knowledge no work to-date has used regulatory loan-level data on the stock to project forward the future path of the entire owner-occupier mortgage market. Second, we use a series of novel techniques and algorithms from Computer Science and Applied Statistics. For example, to estimate the current stock we need to identify previous loans belonging to a borrower moving homes. The challenge arises from the fact that neither borrowers nor properties are identifiable, and from the partial information available, there can be multiple possible matches. To address this, we find the best possible loan matches by maximizing a global scoring function (known as weighted bipartite graph matching (Cormen et al. (2009)). We modify the standard algorithms to make it more scalable with big datasets. While matching techniques have been used in the context of labour economics (Roth (1984)) and housing market related research (Shapley and Scarf (1974)), our work furthers that research direction by applying related algorithms at larger scale for the mortgage market. Separately, we innovatively utilise a statistical approach typically used for risk modelling to take account of the heterogeneous income shocks that households tend to experience.[2]

Our estimate provides an alternative source to track the tail of vulnerable borrowers in the UK. It suggests a larger tail compared to the available household surveys[3]. While the Bank of England/NMG survey points to a falling tail of high DSR loans in recent years, our estimations indicate that it remained almost flat. Similarly, our estimations suggest a consistently higher share of high LTI loans over time. These results should make policy makers less sanguine about the developments in the UK mortgage

[1]Michelangeli and Pietrunti (2014) from Bank of Italy run a microsimulation model combining household-level survey data with macro forecasts on debt and income to project forward the future paths of household indebtedness and DSR. Gross and Poblacin (2016) from the European Central Bank use a macro-micro approach to model the effect of LTV and debt servicing-to-income (DSTI) caps on household sector loss rates, combining GVARs to model house prices, interest rates, stock prices and aggregate unemployment with a household-level logistic model for employment status based on survey data. Cussen et al. (2015) at the Bank of Ireland use regulatory loan-level data from the flow of new loans to assess the impact of LTV measures.

[2]We use GAMLSS, a statistical package (Stasinopoulos and Rigby (2007)) normally used for risk modelling (Tong et al. (2013)).

[3]List of surveys is provided in Appendix 6.8.2

market in recent years, which are traditionally analysed using these surveys. We build a case for relying on these results, both because as they are based on very granular regulatory data and we can validate their robustness against available data. Finally, to illustrate the type of analysis that can be done with the loan-level data, we present the characteristics of high LTV borrowers and movers compared to re-mortgagors.

The rest of the paper is structured as follows. Section 2 explains our approach to estimate the stock of mortgages at the level of individual loans using the loan-level flow data. Section 3 describes the validation of our estimate. Section 4 presents results and Section 5 concludes. More detailed explanation on datasets, the estimation approach and extensions are presented in the Appendix.

2 Estimation Approach

We process the data on the flow of mortgages collected by the FCA to get snapshots of the stock at different points in time. The dataset, Product Sales Data 001 (PSD 001)[4], includes loans for house purchases and re-mortgages since 2005. It has rich information on loan characteristics (including loan amount, interest rate type, repayment mortgage vs interest-only mortgage), borrower characteristics (including date of birth, income, employment status, first-time-buyer vs home-mover) and also property characteristics (including full postcode, type of building, number of bedrooms).[5]

Estimating the stock using the flow data is done in two steps: in a first step we identify mortgages from the flow that are likely to be in the stock at a given point in time, and in a second step we update key borrower and loan characteristics, all of which are collected at the point of origination.

Since 2015, the FCA has been collecting semi-annual snapshots of the stock at the loan-level.[6] Having these data provides us an opportunity to compare our 2015 stock estimation against these reported data to validate our modelling approach. After validating our approach, we repeat this two-step process to produce other point-in-time

[4]The PSD include regulated mortgage contracts only and excludes products such as second-charge lending, commercial, and buy-to-let mortgages. It is commonly referred to as PSD 001 or FCA flow. More information on the dataset can be found here: https://www.fca.org.uk/firms/gabriel/product-sales-data-item.

[5]Despite being a rich dataset, not all variables have been reported for all loans. In Appendix 6.1 we discuss how we identify and remove duplicate loans and impute missing values for key variables. Descriptive statistics on the flow data are also presented in Appendix 6.1.

[6]It is referred as PSD 007 - Mortgage Performance Data (also referred to as PSD 007 or FCA stock data). This dataset tracks the loan performance of each and every existing loan. It captures point-in-time outstanding loan amount, interest rate type, whether the loan is interest-only or repayment, as well as a number of performance metrics such as arrears amount if any.

historical estimates of the stock between 2012-2014.[7].

Below we describe in detail how we construct the stock for 2015[8]. Figure 1 summarizes the stages of our estimation procedure. To estimate the stock of mortgages at the loan-level as of 2015, we start with compiling the loan-level flow data from 2005 to 2015. The first stage is to identify which of the loans that originated after 2005 still exist in the stock in 2015.[9] Not all loans in the combined flow data will be in 2015 stock data. Some of them will have matured or defaulted. Furthermore, the flow data consist of loans to first-time buyers, home-movers and re-mortgagors. The combined flow data therefore contains information on home-movers' and re-mortgagors' both previous and latest mortgages. To avoid double counting, we need to remove home-movers' and re-mortgagors' previous loans. After removing these loans and also the ones defaulted or matured, we are left with the sample of mortgages that are likely to be in the stock as of 2015. The second stage of our estimation is to project loan information, only available at the point of origination, to 2015. Given policy makers and mortgage lenders typically rely on LTI, LTV and DSR[10], we project outstanding loan amount, property value, and income from origination to 2015 so that we can calculate these metrics by 2015 from the stock data.

Below we set out in detail how we identify loans from the combined flow data that are likely be in the stock as of 2015 and update loan characteristics only available at origination. In the next section we present the results from our validation exercises.

2.1 Identifying and removing loans that are not likely to be in the stock

The combined flow data from 2005-2015 contain over 14 million loans while the 2015 FCA stock dataset comprises of only 8 million loans. Part of the discrepancy arises from double counting. A borrower who takes out a mortgage and subsequently re-mortgages generates two separate loan entries in the flow data, but only the re-mortgage appears in the stock data. Similarly, borrowers who take out a mortgage on their new properties when they move homes and pay off the old mortgage loans also appear in the flow twice but only once in the stock. If some borrowers re-mortgage or move home multiple times

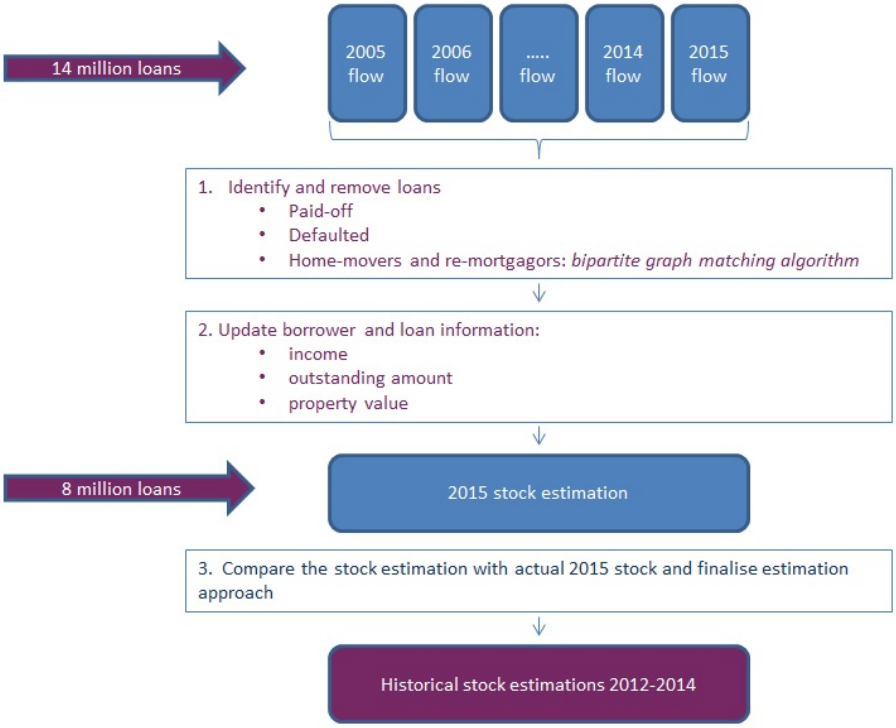
[7]We can estimate the stock for the period before 2012 using the same approach but that may not be as reliable given we have less historical information to estimate the stock for those periods.

[8]When we say 2015 estimate, we mean estimate for 2015 Q2. As that is the time when the FCA stock data started to become available, we can directly compare our 2015 stock estimation against it.

[9]Due to data limitations, our estimate omits loans that originated before 2005Q2. According to the FCA stock data 9 % percent of loans were originated before that.

[10]LTI is a good measure of the financial burden of the borrower. Percentage of borrower's yearly income that goes towards servicing of debts (DSR) measures the affordability as it also takes care of loan interest rate unlike LTI. LTV indicates how much leveraged the borrower is. Hence, all these three indicators together help the lenders and policymakers to understand the quality of individual mortgage.

Figure 1: Estimation Approach



during our period of analysis, their loans may appear even more than twice while only the last mortgage is in the stock. In addition, there are loans in the flow that are no longer in the stock because they will have either defaulted or matured. This section outlines the techniques we use to identify and remove loans from the flow data that are not likely to be in the stock by 2015.

- *paid-off at maturity*: Each loan reports a maturity date at origination. We remove loans which are expected to have been paid off by 2015.[11] This removes around 1.5% loans from the combined flow dataset.

- *defaulted*: Removing defaulted loans is less straightforward. Historical aggregate data indicate the number of mortgages that were possessed each year, but there is no information on which loans defaulted or their characteristics. Evidence [Lambrecht et al. (2003)] suggests that in the UK borrowers tend to default on their mortgages last, only when they run into cash flow constraints.[12] Consistent

[11]We assume that borrowers do not make over-payments, therefore they do not pay off their mortgages before maturity. More details on our assumptions are discussed when have described outstanding loan amount calculation in section 2.2.

[12]Since mortgage debt in the UK is full recourse, we do not see strategic defaults in the UK. On the other hand, in the US when the equity position of a home worsens or becomes negative a mortgagor has an increasing incentive to default on their debt and walk away.

with this, Aron and Muellbauer (2010) find that current DSR is a significant predictor of default. There is also evidence that LTV at origination is a good proxy for secondary factors that are likely to influence probability of default, like savings and wealth [Whitley et al. (2004)]. We therefore use a simple performance matrix based on current DSR and LTV at origination to assign a probability of default to each loan and remove the mortgages that are estimated to perform the worst, calibrating the number of mortgages to remove using the aggregate data. Details on our approach are set out in Appendix 6.2. Around 0.8% of loans are removed from the flow dataset using this approach.

- *home-movers*: Ideally we want to identify each home-mover's previous mortgage from the combined flow data and remove it. The challenge however is that the flow data do not have unique borrower identifiers. Using a home-mover's date of birth we can find all loans that potentially belong to the same borrower ("loan pairs"). However, most of the time there are multiple loans with the same date of birth. This means that the same borrower has moved home multiple times or that there are different borrowers with same date of birth[13]. We need a way to identify the loan pairs that most likely to belong to the same borrower. First, we track "mover chains": these are chains of three transactions, two belonging to the borrower moving homes (for the original property and the new one they are moving to) and one to another borrower purchasing the home-mover's old property. We can link these transactions given that the borrower moving into the home-mover's old property generates a loan with the mover's old mortgage postcode, and be taken out within a few days of the mover's new mortgage [Figure 2]. This allows us to effectively triangulate transactions with more certainty. As there are still possible multiple matches, we use a weighted bi-partite graph-matching algorithm to identify all the loan chains and select the ones that are most likely to be the real chains in a way that is globally optimal (i.e., overall matching quality among all the possible chains is optimised rather than just picking the best matching for individual home-mover in some order). Home-movers' previous loans identified in this way are removed as a first step. We can track around 60% of recent (2014-15) home-movers' previous mortgages in this way[14].

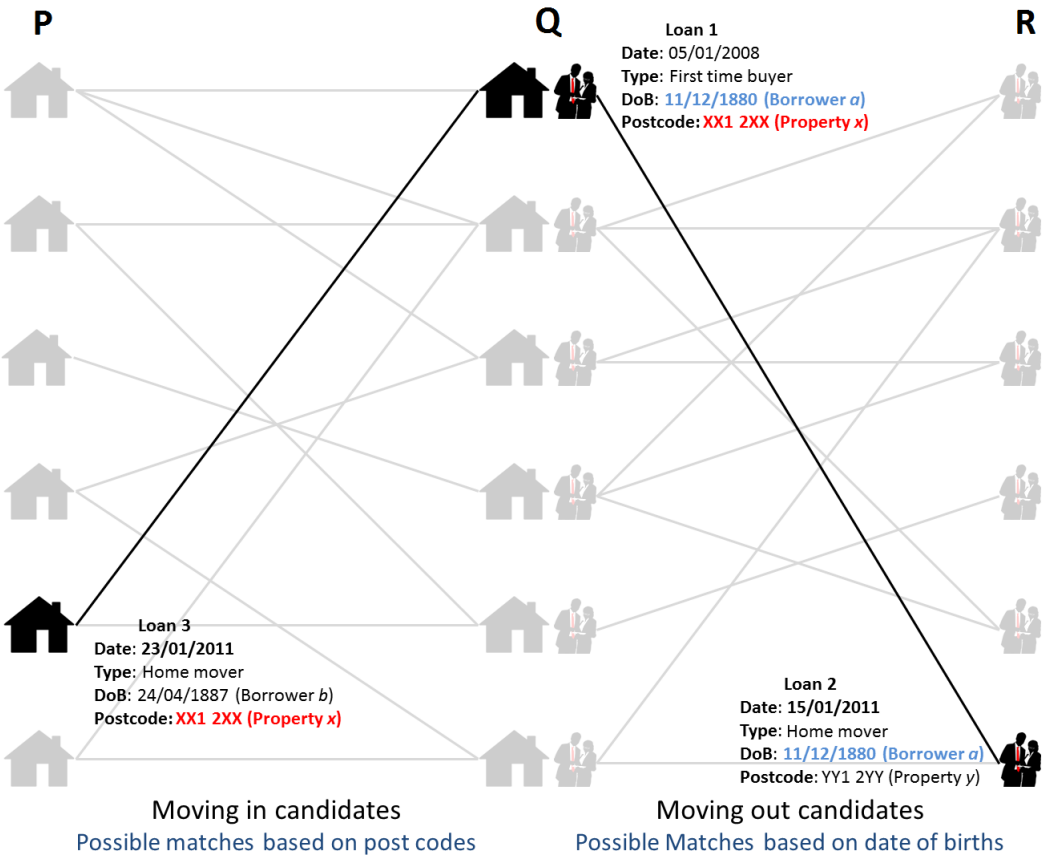
  However, this method does not work if a chain in the link is missing; for example, if there is a last-time seller, cash-buyer or buy-to-let investor among the counter-parties then the transaction is not captured in the flow data. Therefore, for the remaining transactions, we design and implement a similar weighted bi-partite graph-matching algorithm to identify all loan pairs most likely to belong to the

[13]This is highly likely given that there are more than 14 million loans.

[14]Our tracking percentages are less for earlier years as previous mortgages of corresponding home-movers have higher probability to be prior to 2005 which we do not have in the flow data. However, this is not a problem as it means we do not have any previous loans to remove for these home-movers from our flow while estimating the stock.

same borrower, again in a way that is globally optimal. The algorithms work by maximising a global objective function based on a score that captures the "match quality" for each loan pair based on the proximity of the two homes in terms of geographical location and house price, among others. We can track additional 35% of recent (2014-15) home-movers' previous mortgages approximately in this way with lower tracking percentages for older years as expected. More detail can be found in Appendix 6.3. Around 18% of the total loans are identified as home-movers' previous loans in these two steps and removed for the 2015 stock.

Figure 2: Illustrative mover chains



- *re-mortgagors*: We identify the previous loan for each re-mortgage using property post-codes and borrowers' date of birth. Where there are multiple matches, the best candidate is picked based on quantitative and qualitative characteristics, including change in income and property characteristics where available. We can track around 51% of recent (2014-15) re-mortgagors' previous mortgages in this way. We track fewer re-mortgages in earlier years as the flow data start in 2005, and older original loans will not be in the stock anyways. There are fewer cases where multiple potential matches arise compared to home-movers, so we do not

apply a global optimisation function at this stage. Instead, the best loan pairs are chosen sequentially by the date of origination. This allows us to identify and remove 13% of mortgages for 2015 stock estimation.

We also attempt to identify loan pairs where the date of birth may have changed during the re-mortgage. This may happen for loans that are jointly owned where the borrowers or lenders may register one date of birth for an initial loan and then the other's date of birth in the re-mortgage. In such cases we rely on post-codes and a number of conditions to capture possible candidate loans for each re-mortgage. For example, we arbitrarily restrict the age difference of the borrowers to 10 years, constrain the change in property value and rely on qualitative information about the property (for more detail see Appendix 6.4). In this second step, we use a weighted bi-partite graph-matching algorithm similar to one that we use to identify home-movers. We can track additional 36% recent (2014-15) re-mortgagors in this way with lower tracking percentages for earlier years as expected. We identify and remove an additional 8% mortgages for 2015 stock estimation.

2.2 Projecting variables of interest forward

Once we have identified which loans are likely to be in the stock, the next step is to project the three key variables that allow us to calculate LTI, DSR and LTV ratios for each borrower as of 2015: borrowers' incomes, outstanding loan amounts and property values.

- *Borrower income projection.* A crude approach to project incomes from origination to 2015 would be to use aggregate statistics, such as household disposable income growth, and update every borrower's income in line with average income growth. We can use more granular statistics, such as local authority data on average earnings. The main limitation with these methods is that they omit heterogeneous income shocks borrowers might experience after taking out a mortgage, as described by Anderson et al. (2014). These are particularly important for estimating the tail of highly indebted households, as evidence suggests these borrowers often end up with high DSRs as a result of negative shocks. To address this, we simulate heterogeneous income shocks conditional on the borrower's employment status, age and income level. We find that self-employed borrowers have more volatile income and income variability depends on whether they are high earners or low earners to begin with. Moreover, borrower income can change significantly when they reach retirement age. The model is calibrated using empirical data from two household-level panel surveys: *British Household Panel Survey* (BHPS) and *Understanding Society* (USoC). The modelling approach is explained in detail in Appendix 6.5.

- *Outstanding loan amount projection.* The calculation of outstanding loan amounts

by 2015 is based on several assumptions. There are three types of mortgage repayment schedules: (i) capital and interest, (ii) interest-only, and (iii) mixed. Mixed mortgages are composed of capital and interest, and interest-only components. We do not have information on the share of these components, so we assume that mixed type mortgages are composed of 50 percent capital and interest, and 50 percent interest-only components.

  - For interest-only mortgages (and the interest-only component of mixed type mortgages), outstanding balances are kept the same as loan value at origination since capital is only paid at maturity.
  - For capital and interest mortgages (and the capital and interest component of mixed type mortgages), outstanding amounts as of 2015 are calculated by taking into account monthly payments for the period from origination of each loan to 2015.

  For monthly payment calculations, we adjust interest rates depending on the type i.e. fixed, flexible, tracker or discount. The details of the outstanding amount calculation and underlying assumptions are explained in Appendix 6.6. We do not adjust for under- or over-payments given lack of available data to calibrate such as how many borrowers are missing payments or making payments outside of schedule as well as which borrowers are over/under paying and by how much.

- *Property value projection.* Property values at origination are updated in line with house price growth at different geographical granularities using Land Registry Price Paid data and officially published regional house price indices. Land Registry provides information on every house purchase in England and Wales since 1995. Using these data, house price growth is calculated for all possible year-quarter pairs between 2005-2015 at NUTS (Nomenclature of Territorial Units for Statistics European Commission (2017)) and county level. Property values in the flow data are updated using the median value of the granular house price growth estimated using Land Registry and the official regional statistic. Although more granular geographical region index can capture regional dynamics, these indices are calculated by us based on sale prices of transactions that take place without property characteristic adjustment. Properties are generally sold when their prices are sufficiently appreciated, so indices based on transactions tend to be biased upwards. Also, smaller geographic areas have fewer properties sold per quarter making the indices more volatile. By contrast, the UK regional index, while less granular, is adjusted for some of these biases. More details can be found in Appendix 6.7.

3 Validation

Validating our approach is crucial, in particular given it is based on several assumptions. We check and compare our estimate against the FCA stock data[15] and two household surveys[16]: the Wealth and Asset Survey (WAS) and the Bank of England / NMG Survey (NMG). We also supplement with aggregate statistics where possible.

We start by validating the first stage of our estimation: check whether we select a representative subset of borrowers for the 2015 estimate. Figure 3a and 3b show the LTV and LTI distributions at origination for both of our estimation and the FCA stock data as of 2015 are comparable. Figure 3c and 3d provide a comparison of age distribution and employment status between our estimation, and the FCA's stock data and WAS. These results indicate that borrowers selected for the stock estimate are indeed representative.

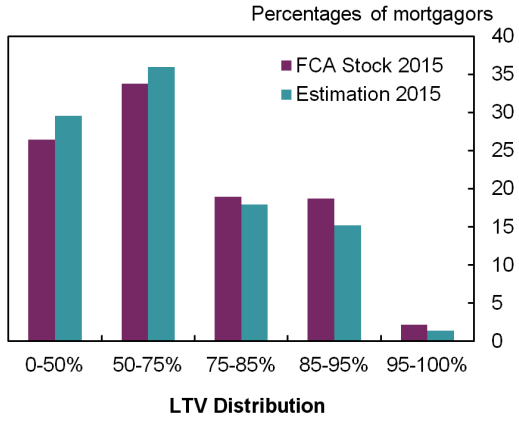
Next we look at the second stage of our estimation by comparing our projections of outstanding loan amounts, borrowers' incomes and property values in 2015 with the FCA stock data and household surveys. Relying on survey data for validation presents challenges. For example, Anderson et al. (2016) document that household surveys tend to underestimate mortgage debt. This might be due to survey participants under-reporting their outstanding debt or under-representation of high-debt mortgagors in surveys. The FCA stock dataset is more reliable, as it is based on the entire population of mortgages and is not subject to reporting bias. The main limitation however is that it does not capture updated incomes, which are central to our estimate for pinning down vulnerable borrowers.

As shown in Figure 4 and Table 1, while household surveys such as WAS and NMG underestimate the average loan size, our projection performs better, matching statistics from the FCA stock data and ONS aggregate data more closely. Consistent with this, Figure 5a shows the distributions of outstanding loan amount for both the FCA stock and our projection are shifted towards higher loan values compared to surveys. Notably, our estimate closely matches the FCA data, which takes into account under- and over-payments. This assuages potential concerns from the fact that we do not adjust the loan amount for these.

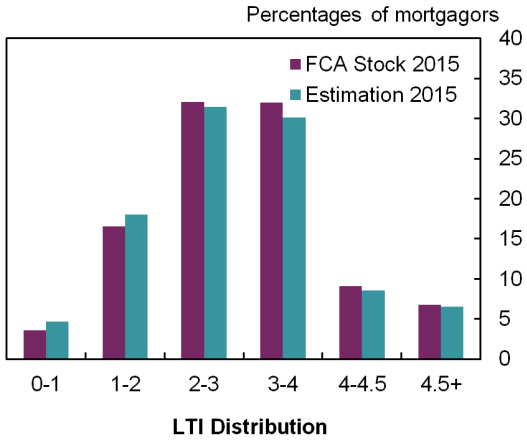
Figure 5 shows income distribution from our 2015 income projection against the WAS, the most reliable household survey on income. One caveat in this comparison is

[15]We have loan-level stock information for 2015 from the FCA, which provides an opportunity to compare our estimation against the stock population. In order to conduct this comparison we need to merge the FCA stock data with PSD 001 as many of the useful variables are missing in the FCA stock data. There is no straightforward way to merge these two datasets, so we used matching criteria as described in Appendix 6.8.1.

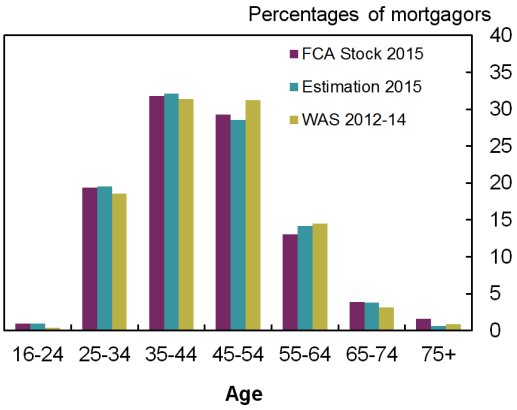
[16]Details of different surveys we have used in our work are mentioned in Appendix 6.8.2



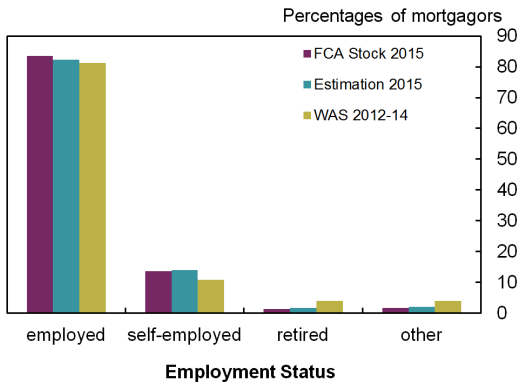
(a) Loan-to-Value distribution at origination



(b) Loan-to-Income distribution at origination



(c) Distribution by age



(d) Distribution by employment status

Figure 3: Comparing our estimated stock with FCA stock and WAS

that while our projection is for 2015, the latest available income data from WAS spans from 2012-2014. They look comparable, though our projection suggests a higher share of borrowers in the highest income buckets (above £100,000).

Figure 5c shows a discrepancy in the distribution of property values between our projection and the WAS. This might be due to bias in household surveys as households may not be able to estimate accurately the current value of their properties. Besides, our projection (2015) and the survey data (2012-2014) are for different time periods. On the other hand, our property value projection might be biased as well since it is based on the house sale/purchase prices from Land Registry. Since the owners generally sell properties only when the prices have increased and hold on otherwise, our estimate for all house prices might be upwardly biased as it relies on transactions that took place.

Figure 4: Average loan per mortgagor

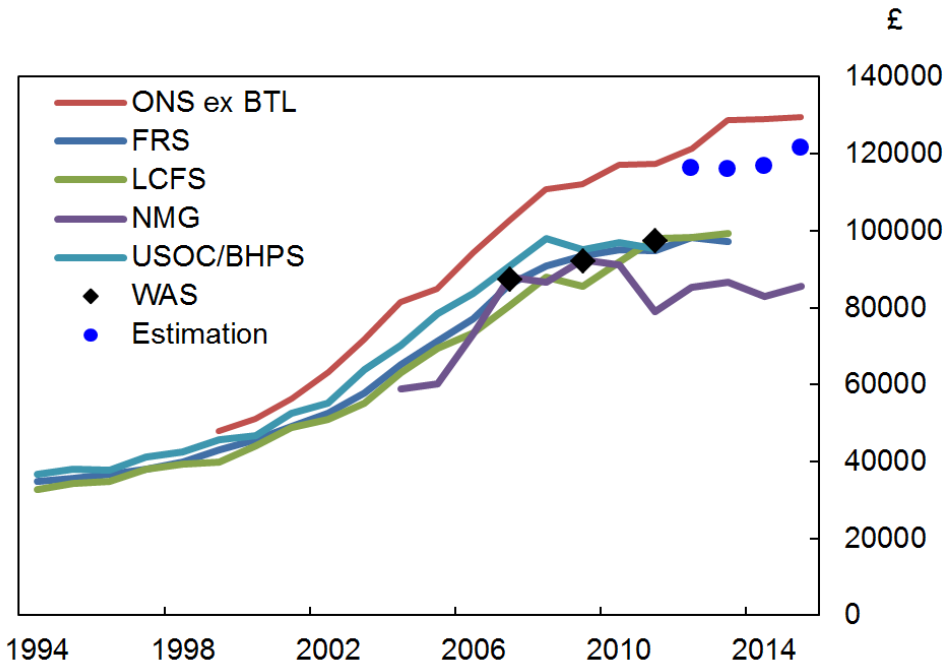
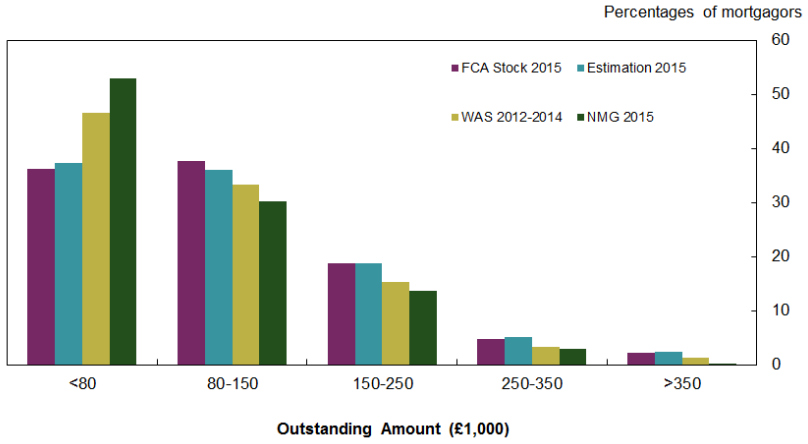
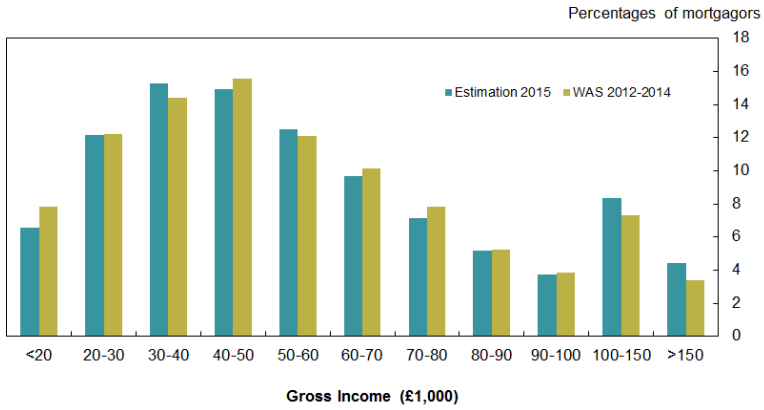


Table 1: Outstanding loan amount comparison

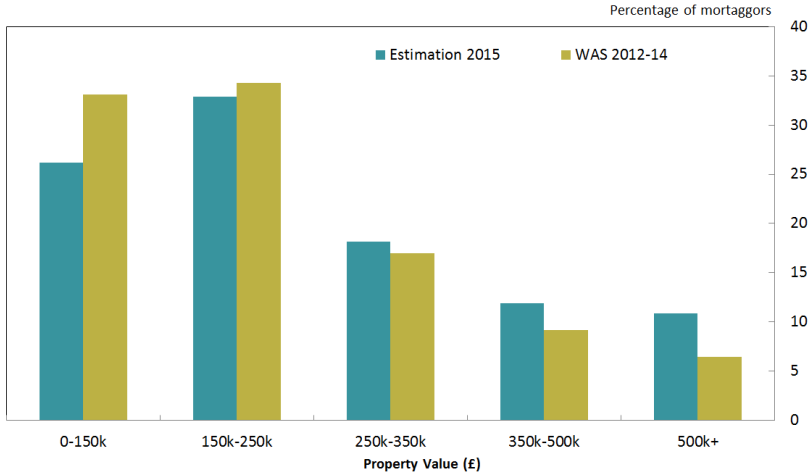
Sources	Average loan
Estimation 2015	£121,518
FCA stock 2015	£124,244
ONS excl. BTL 2015	£129,643
NMG 2015	£87,322
WAS 2012-2014	£98,814



(a) Outstanding loan distribution



(b) Income distribution



(c) Property value distribution

Figure 5: Comparing our projections with FCA stock and surveys

4 Results

4.1 Loan-level stock distributions

In the absence of loan-level stock of mortgages in the previous years, policy makers have been relying on survey data to monitor risks to financial stability and calibrate policies. As already discussed, surveys can be subject to biases and small sample issues, so our work provides an alternative estimate of the LTI, DSR and LTV distributions. Notably, we find that size of the tail of vulnerable borrowers might have been higher in recent years than surveys suggest.

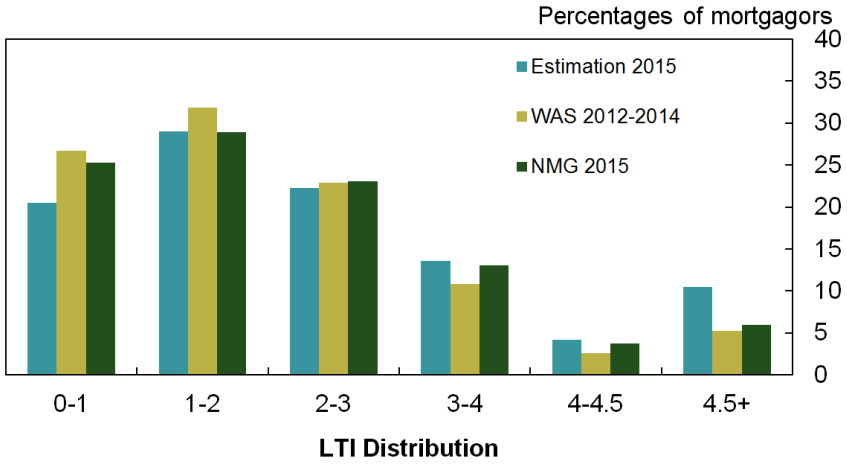
Figure 6 shows LTI, DSR and LTV distributions from our estimation against data from the NMG and WAS surveys as of 2015. Our estimate suggests a larger vulnerable tail (LTI above 4.5 and DSR above 40) compared to surveys. This is consistent with the finding that individuals tend to underestimate outstanding loan amount in surveys or highly indebted borrowers are under-represented in surveys. There is greater discrepancy in LTV distributions, which could be attributed to further bias in how individuals report property values in surveys.

Figure 7 compares the evolution of the tail of high LTI and DSR from our estimate and surveys over time. While the NMG survey suggests that the share of loans with DSR above 40% has decreased in recent years, our estimations indicate that it remained almost flat. Similarly, our estimations for high LTI shares are steadily higher.

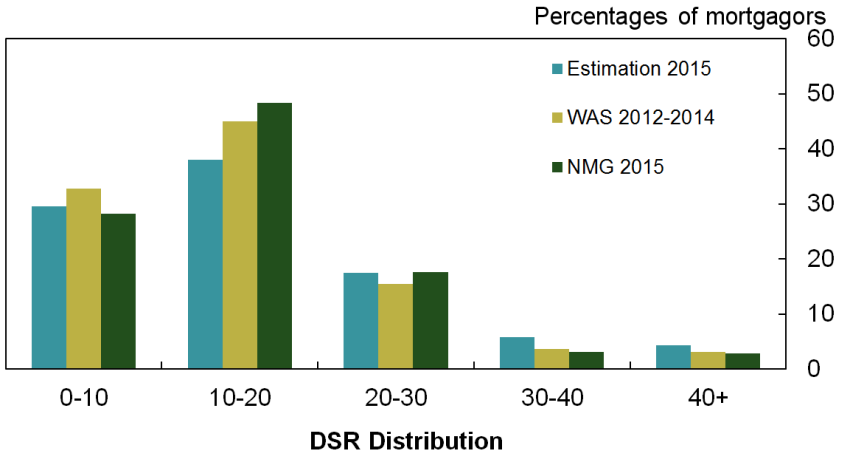
Our stock estimation can also shed light on the characteristics of specific cohorts of borrowers and loans. To illustrate the type of analysis can be done by using the loan-level estimate, we present the age distribution of high DSR (40%+), high LTI (4.5+) and high LTV (85%+) loans.

Figure 8 shows the characteristics of current high LTV loans. We find that most of the LTV loans were originated pre-2009, before credit conditions tightened materially in the UK.[17] We also see that a large proportion of outstanding loans with high LTVs (85%+) are interest-only. This suggests that if policy makers are concerned with the tail of high-LTV mortgages that are interest-only, analysing the flow of new loans is unlikely to provide much insight, as interest-only loans are currently very rare. The focus needs to be turned to the high-LTV cohort in the stock and the interest-only loans originated in pre-crisis period.

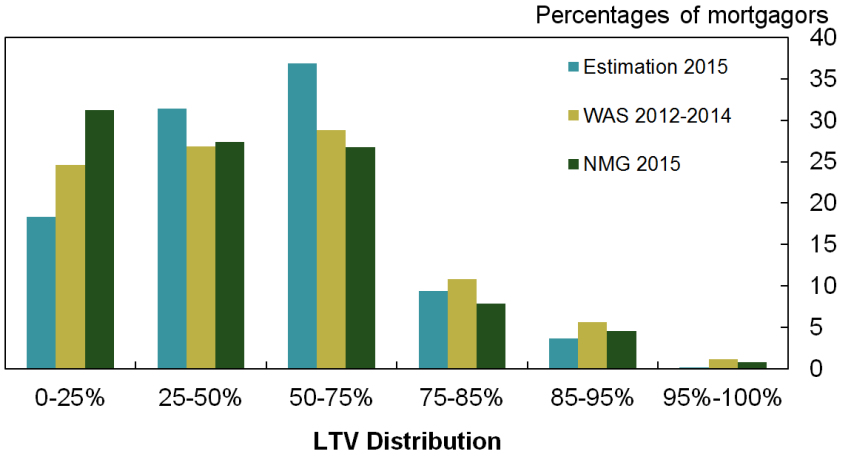
[17] We have excluded 2014-2015 loans from the chart as they are originated very recently and hence their current LTVs have not changed significantly within less than 1.5 years.



(a) Loan-to-Income distribution

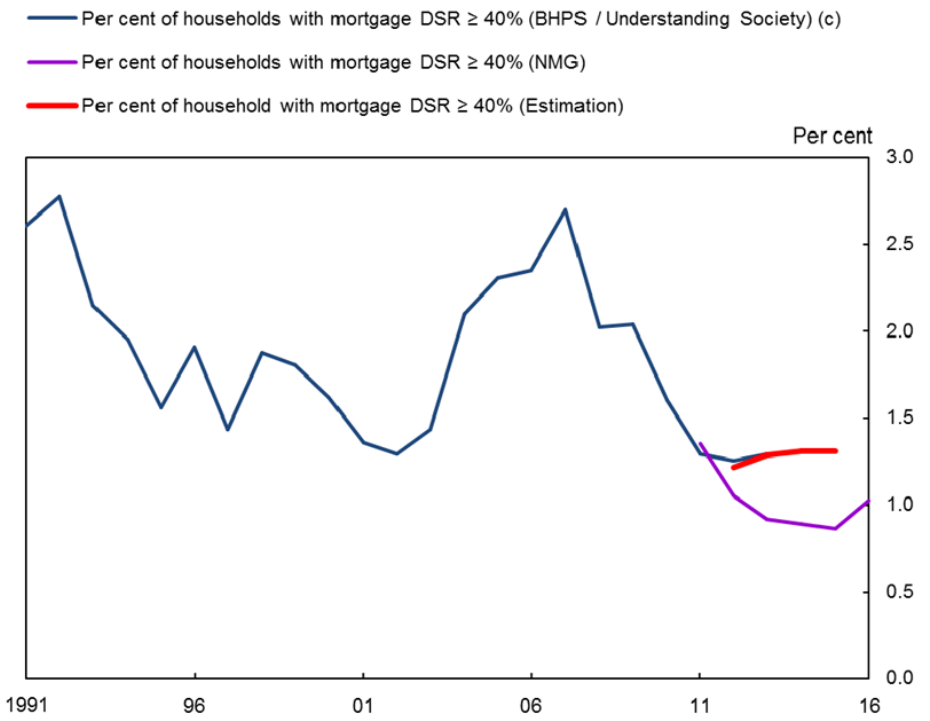


(b) Debt Service Ratio distribution

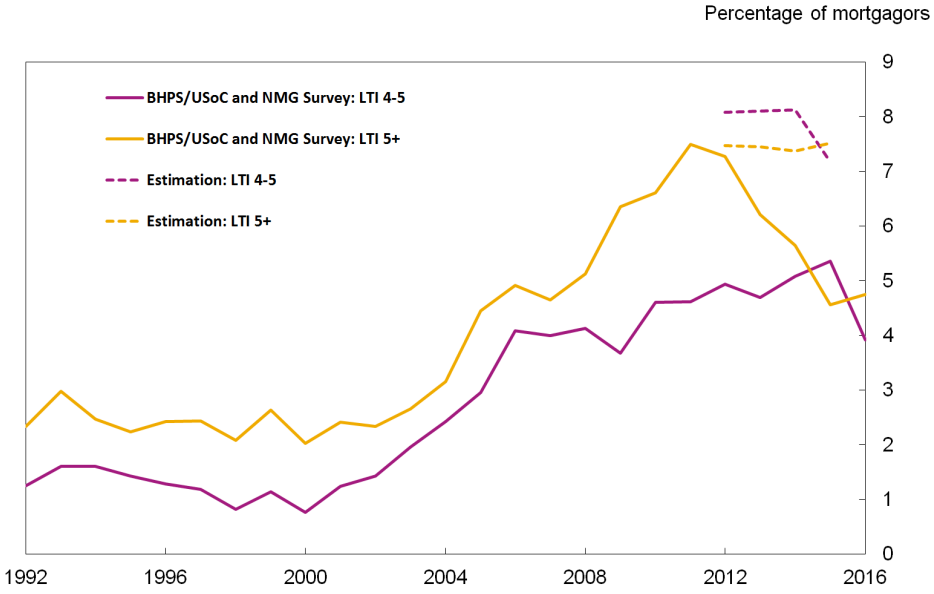


(c) Loan-to-Value distribution

Figure 6: Comparing three main ratios



(a) Historic estimation of DSR with survey results



(b) Historic estimation of LTI with survey results

Figure 7: Comparing the ratios with surveys

Figure 8: Current high LTV loan share by type of loan repayment and year of origination

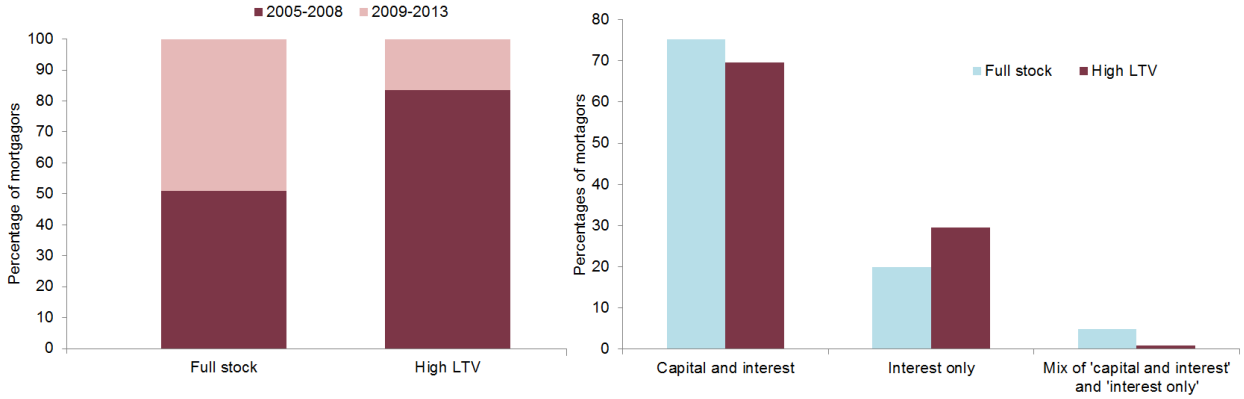
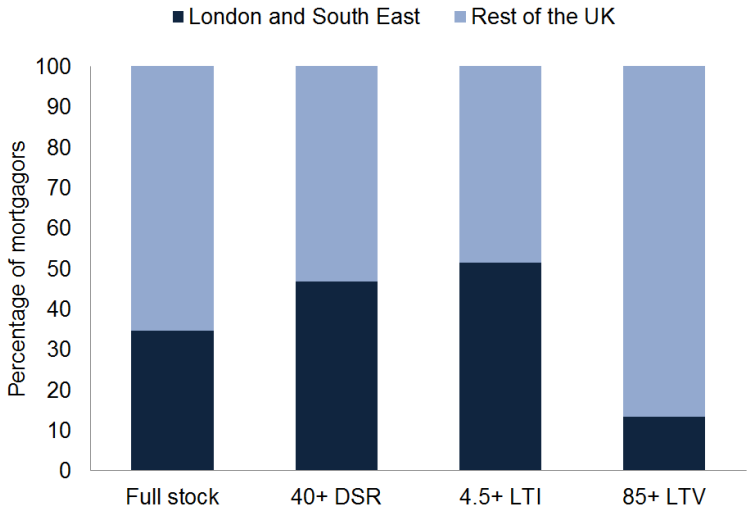


Figure 9: Regional concentration of loans by different type of tails



We can also analyse the geographical distribution of high DSR, high LTI and high LTV loans. Figure 9 shows that high DSR and LTI loans are concentrated in London and the South East of England, while high LTV loans tend to be in the rest of the UK. This is consistent with borrowers in London and South East benefiting from higher house price growth, lowering their current LTVs. Equally, as house prices tend to be higher, these borrowers are also more leveraged at origination.

In the process of estimating the stock we also uncover interesting insights into the dynamics of the mortgage market. From matching the FCA stock with the flow data (matching techniques are discussed in Appendix 6.8.1) we can estimate the share of

borrowers switching their mortgage repayment type and interest-rate type respectively on their existing loan (Table 2 and 3). The results suggest that capital and interest mortgagors do not tend to switch much, whereas borrowers with mixed repayment type mortgage are more likely to switch to capital and interest repayment schedule. Not surprisingly, mortgagors tend to switch to standard variable rate deals (SVR).

Table 2: Repayment type in 2015 H1 as a percentage of different repayment types at origination

	2015 H1 repayment type		
Repayment type at origination	**Capital and interest**	**Interest only**	**Mix**
Capital and interest	97 %	2 %	1 %
Interest only	6 %	81 %	13 %
Mix	21 %	7 %	72 %

Table 3: Rate type in 2015 H1 as a percentage of different rate types at origination

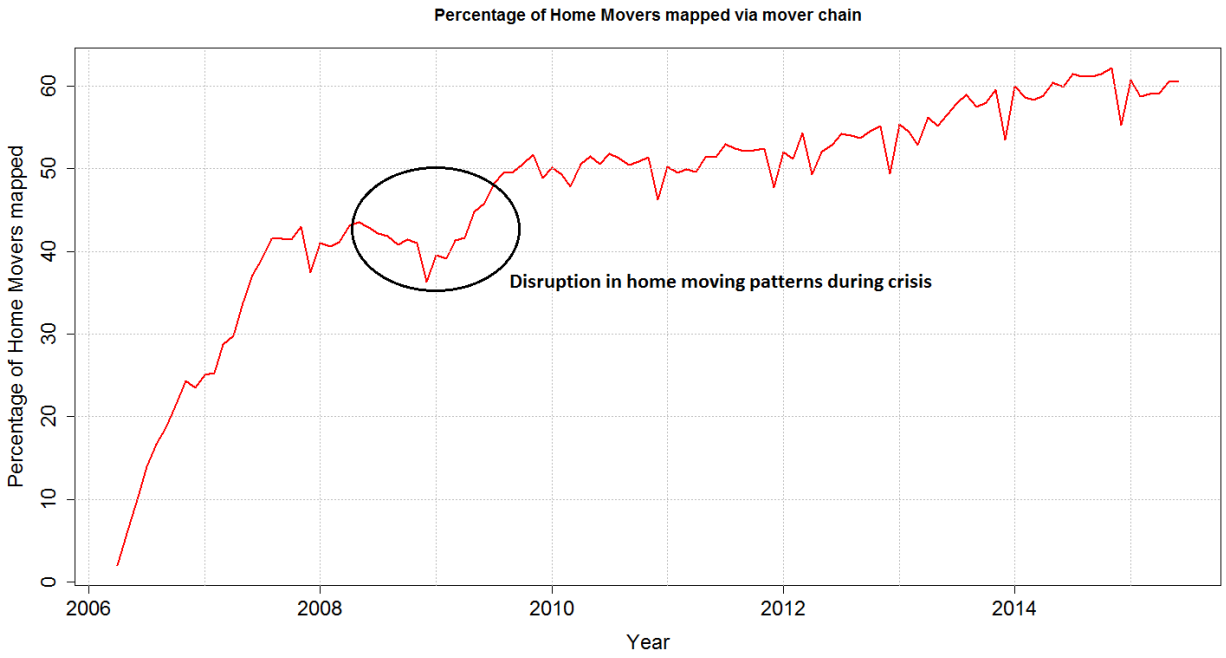
	2015 H1 rate type			
Mortgage rate type at origination	**Fixed**	**Tracker**	**SVR**	**Other**
Fixed	59 %	9 %	29 %	3 %
Trackers	14 %	55 %	25 %	6 %
SVR	7 %	9 %	36 %	47 %
Other	20 %	41 %	28 %	12 %

4.2 Analysis on home-movers and re-mortgagors

In identifying movers and re-mortgagors, we uncover interesting patterns that shed light on borrowers behaviour. Figure 10 shows the percentage of home-mover mortgages we could track using the "chains" method over time (i.e. the share of home-movers in each year for whom we could trace their previous loan). The proportion rises rapidly in the first few years, as the number of historical flows containing more candidates for matching become available from 2005. The share plateaus at around 60%, likely because some of the links in the mover-chains are not mortgaged and therefore will not be captured by this method.

Interestingly there is a fall in the share of movers we can track around 2008-09 during the global financial crisis. We believe that the break in the trend reflects a temporary shift in the mix of borrowers moving homes rather than an underlying feature of our algorithm to identify movers. In particular, fewer borrowers who made a purchase immediately pre-crisis (and whom we can track from the flow post 2005) were moving during this period. This could be because borrowers who bought their homes during

Figure 10: Tracking home-movers via mover-chains



the boom immediately before the crisis were more likely to have found themselves "underwater" when house prices fell, limiting their ability to move. The shift appears to normalize as house prices started to recover and mortgage rates began falling, allowing borrowers who bought just before the crisis to come out of negative equity and move homes. This illustrates how identifying transaction chains can uncover interesting patterns around home-mover behaviour (more details are in Carvalho et al. (2016)).

Identifying movers and re-mortgagors also allows us to track certain borrower characteristics which can help us better understand how households make their mortgage decisions. Figure 11 shows the evolution of borrower income, based on their reported income in the initial loan and at the point of either move or re-mortgage. The first chart shows the distribution of annual income growth between these two transactions. The longer tails point to home-movers experiencing more extreme income shocks relative to re-mortgagors. The second chart shows how the median annual income growth for borrowers moving or re-mortgaging in a given quarter. Home-movers consistently experience higher average income growth relative to re-mortgagors. Both of these facts are consistent with borrowers choosing either to down-size or trade-up following a large change in affordability.

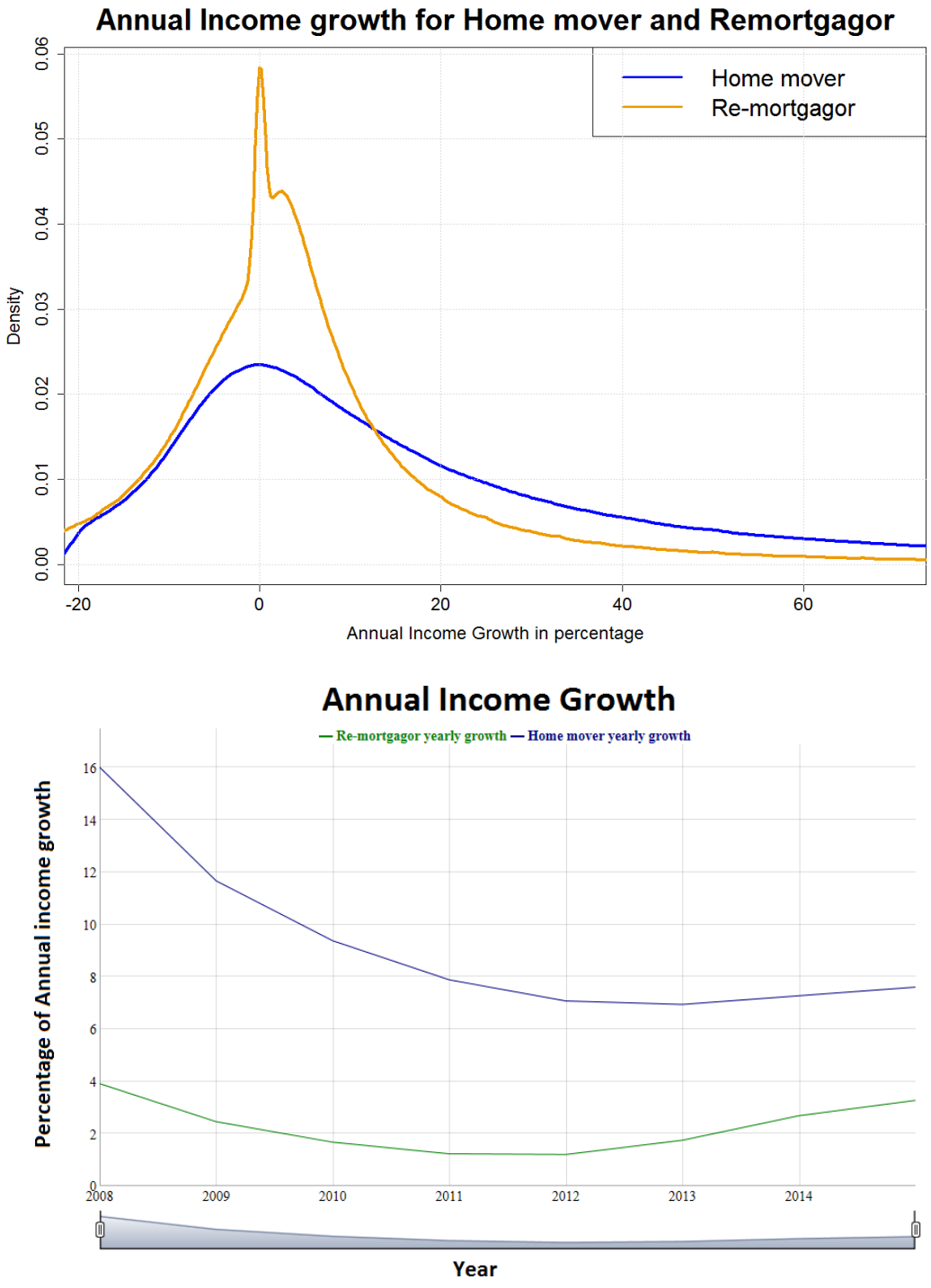


Figure 11: Comparing re-mortgagors and home-movers

5 Conclusion

We have developed one of the first approaches to estimate the characteristics of the stock of mortgages at the loan-level from the flow data by using the techniques and algorithms normally used outside the domain of Economics.

Our estimate provides an alternative source to track specific cohorts of borrowers in the UK mortgage market. We find that a larger tail of vulnerable borrowers than household surveys suggest. While survey data suggest that the share of high DSR loans has decreased in recent years, our estimations indicate that it remained almost flat. Similarly, our estimate of high LTI shares over time are steadily higher than surveys. All these results suggest that policy makers should be less sanguine about the developments in the UK mortgage market in recent years. As these analyses are based on very detailed granular regulatory data, we believe the results are more reliable than surveys.

In addition, we uncover several interesting patterns for home-movers and re-mortgagors and find quantitative evidence on how the patterns have evolved by cohorts, type of loans and housing choice. We also present the characteristics of high LTV borrowers and compare movers against re-mortgagors to illustrate the type of analysis that can be done with the loan-level data. Having a disaggregated estimate of the stock can help us better understand different segments of the market, and also catalyse external academics to collaborate with the Bank for further research projects.

Our research also forms a foundation to model how the characteristics of the stock of loans may evolve under different macroeconomic scenarios. It allows policymakers to use loan-level stock data to simulate various policy proposals and better predict how those proposals might affect the UK mortgage market, and in turn, the outlook for financial stability.

There is scope for further extending the analytical side of the work. More advanced predictive models, in particular supervised machine learning algorithms (such as neural networks or support vector machine) could be implemented to estimate under and over-payments and possessions. The objective functions we implement to track the re-mortgagors and home-movers can also be made more robust to capture more subtle borrowers' behaviour. Further, as the data volume grows over time, we can afford to use more sophisticated techniques as it ensures the noise in the data does not pollute the results. Finally, as the big data tools are also becoming more powerful we expect our work will feed into further research on such granular regulatory data in the future.

Bibliography

Anderson, G., Bunn, P., Pugh, A., and Uluc, A. (2014). The potential impact of higher interest rates on the household sector: evidence from the 2014 NMG consulting survey. *Bank of England Quarterly Bulletin*, 54(4):419–433.

Anderson, G., Bunn, P., Pugh, A., and Uluc, A. (2016). The Bank of England/NMG Survey of household finances. *Fiscal Studies*, 37(1):131–152.

Aron, J. and Muellbauer, J. (2010). Modelling and forecasting UK mortgage arrears and possessions. *Department of Economics Discussion Paper Series, University of Oxford*, page Ref: 499.

Bank of England (2013). Short-term risks to financial stability. *Financial Stability Report*, page 25.

Bank of England (2014). Box 6: International experience wtih macroprudential mortgage product instruments. *Financial Stability Report*, page 63.

Batista, G. E. A. P. A. and Monard, M. C. (2002). *A Study of K-Nearest Neighbour as an Imputation Method*. IOS Press.

BBC (2004). Self-cert mortgages could skew market. `http://news.bbc.co.uk/1/hi/business/3478635.stm`, last accessed: 21 August 2017.

Bunn, P. and Rostom, M. (2015). Household debt and spending in the United Kingdom. *Bank of England Staff Working Paper No 554*.

Carvalho, A., Chakraborty, C., and Latsi, G. (2016). 'Matchmaker, matchmaker make me a mortgage': What home-movers can learn from dating websites. *Bank Underground, Bank of England staff Blog*. `https://bankunderground.co.uk/2016/07/27/matchmaker-matchmaker-make-me-a-mortgage-what-policymakers-can-learn-from-dating-websites/`.

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2009). *Introduction to Algorithms (Third ed.)*. MIT Press and McGrawHill. ISBN: 978-0-262-03384-8, Pages: 732-759.

Cussen, M., O'Brien, M., Onorante, L., and O'Reilly, G. (2015). Assessing the impact of macroprudential measures. *Economic Letters 03/EL/15, Central Bank of Ireland*.

Daniel, W. W. (1990). Kolmogorov-Smirnov one-sample test. *Applied Nonparametric Statistics (2nd ed.). Boston: PWS-Kent. pp. 319-330*. ISBN: 0-534-91976-6.

European Commission (2017). NUTS - nomenclature of territorial units for statistics. `http://ec.europa.eu/eurostat/web/nuts`, last accessed: 21 August 2017.

Goldberg, A. V. and Tarjan, R. E. (1986). A new approach to the maximum flow problem. *Proceedings of the eighteenth annual ACM symposium on Theory of computing - STOC*. ISBN: 0897911938 doi:10.1145/12130.12144.

Gross, M. and Poblacin, J. (2016). Assessing the efficacy of borrower-based macroprudential policy using an integrated micro-macro model for European households. *ECB Working Paper No. 1881*.

Lambrecht, B. M., Perraudin, W., and Satchell, S. (2003). Mortgage default and possession under recourse: A competing hazards approach. *Journal of Money, Credit and Banking*, 35(3):425–42.

Mian, A., Rao, K., and Sufi, A. (2013). Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*.

Michelangeli, V. and Pietrunti, M. (2014). A microsimulation model to evaluate Italian households financial vulnerability. *Questioni di Economia e Finanza (Occasional Papers) 225, Bank of Italy, Economic Research and International Relations Area*.

Roth, A. E. (1984). The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy*, Volume 92:991–1016.

Shapley, L. and Scarf, H. (1974). On cores and indivisibility. *Journal of Mathematical Economics*, Volume 1:23–28.

Stasinopoulos, M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS). *Journal of Statistical Software*, Volume 23(7).

Tong, E., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, Volume 29:548562.

Whitley, J., Windram, R., and Cox, P. (2004). An empirical model of household arrears. Bank of England Staff Working Paper no. 214.

6 Appendix

6.1 FCA flow data preparation

6.1.1 Removing duplicates and aggregating multiple loan entries

In the flow, 3% of the loans have the exact same date of origination, property value, postcode, date of birth and gross income of the borrower and either same or slightly different loan values. There could be two explanations for these duplicate loans. One possibility is that loan information was entered corresponding to a particular loan but with some incorrect details. Then the correct details were re-entered but the old entry was not removed. The other possibility is that a given loan is split into multiple loans. If that is the case, then we should definitely recognise and consolidate them, otherwise they will have smaller LTVs and LTIs, giving a false understanding of the quality of the mortgage pool. To handle this issue we follow the approach described below:

For the loans which have the exact same date of origination, property value, postcode, date of birth and gross income of the borrower, and

- same loan value: we consider them as duplicate and keep only the latest entry,
- slightly different loan value: we pick the one with highest loan value. However, if the highest loan value is still quite small, we combine multiple loan values into a single one, unless the addition is greater than the property value. When combining multiple loans into single loan, we take the weighted average of the loan characteristics where applicable, such as interest rate where weights correspond to loan value.

6.1.2 Imputing missing values

Although the flow is a rich dataset, not all variables have been reported for all loans. To be able to estimate outstanding loan value and calculate metrics on indebtedness, we impute missing values for key variables. Table 4 summarises the percentages of loans where either the mortgage interest rate, borrower income or loan term were not reported. We calculate these missing values using a k-nearest neighbour technique (Batista and Monard (2002)). This technique first involves finding the k most similar loans to the loan with a missing variable based on a set of criteria. Then a weighted average of the missing variable is taken from those k loans. For example, if a mortgage entry does not have information on borrower income, we search for other loans in the same postcode where the borrowers are of similar age and then take an average of income of the two loans nearest in terms of property value.

Table 4: Loans missing key variables of interest

Total number of loans	14,162,671
Missing interest rate	32%
Missing income	2%
Missing mortgage term	1%

6.1.3 Descriptive Statistics

Table 5: The flow data statistics: Share of different types of loans and median values of certain important variables are mentioned below by the origination

	Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Loan	Fixed rate share (%)	63	64	72	58	67	49	62	68	85	88	88
	Interest only share (%)	26	28	32	31	23	19	15	10	6	5	0
	Mortgage term in years	23	23	23	21	20	21	21	23	24	25	25
	Loan value (1000 £)	98.5	107	115	115	106	112	111.5	116	121.5	131	136.4
	Interest rate in %	5.00	4.99	5.74	5.83	4.39	3.69	3.48	3.69	3.14	3.05	2.69
	Fixed period length (years)	3	2	2	3	3	2	2	2	2	2	2
Property	London/South East (%)	32	33	33	34	35	35	35	36	37	37	36
	Property value (1000 £)	160	170	183	190	180	190	188	189.95	192.5	200	217.5
Borrower	First time buyer share (%)	16	16	16	13	20	22	21	24	27	30	28
	Home-movers share (%)	28	31	31	22	33	38	34	36	35	36	35
	Re-mortgagors share (%)	51	48	48	60	42	35	40	35	34	30	33
	Age in years	38	38	38	39	40	40	40	39	38	37	37
	Verified income share (%)	62	57	50	48	57	64	70	77	82	88	100
	Employed share (%)	79	78	78	80	82	83	84	85	85	85	85
	Retired share (%)	2	2	2	2	3	4	3	3	3	3	1
	Self-employed share (%)	16	17	18	16	12	11	10	10	10	10	10
	Gross income (£)	36,816	38,500	40,384	42,000	40,942	41,285	41,360	42,080	43,522	45,769	48,345

Table 5 shows that the share of interest-only mortgages is decreasing and fixed rate mortgages is increasing. We can also see that all mortgages are now income-verified (as opposed to the pre-crisis period where a material share of borrowers self-verified their income when applying for a mortgage).

6.2 Performance matrix to identify and remove loans that are most likely to be defaulted

We create a simple performance matrix[18] based on DSR as of 2015 and origination LTV using the FCA stock data, which contains information on arrears. We first group the FCA 2015 stock data into multiple buckets based on their current DSR (iDSR) and origination LTV (oLTV.) We consider 30 equal size iDSR buckets and 7 oLTV buckets. For each of those 210 (30×7) buckets, we calculate the percentage of mortgages in arrears for more than 6 months[19] but less than 24 months (this is the definition of default we have used here). After creating the matrix, we translate this to our estimated stock data. We group our data into the same 210 buckets and randomly pick an equivalent percentage of loans and mark them as defaults. For example, suppose the FCA stock data suggest that 1.7% of loans in the 32-34 iDSR and 80-90 oLTV range are in default. We then pick a random 1.7% of loans in the same bucket from our estimated stock data and mark them as defaults. Note that as we move towards higher iDSR and oLTV buckets, default percentages increase.

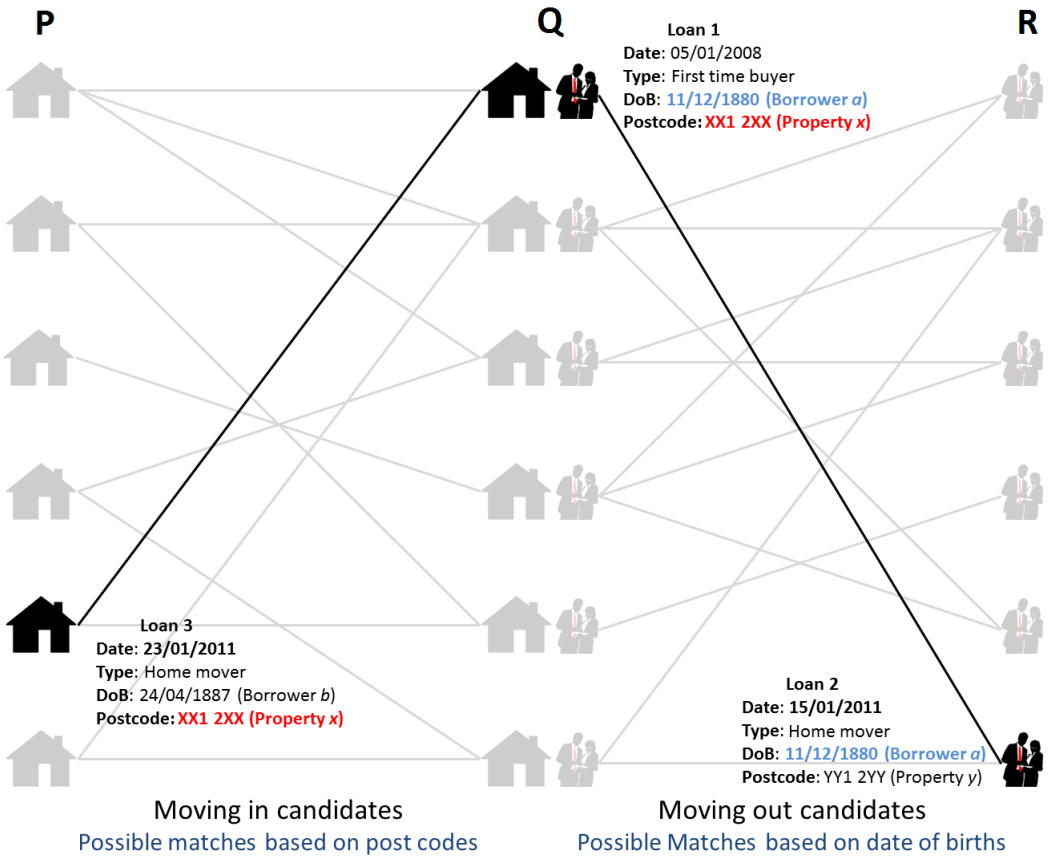
6.3 Identifying previous mortgages of home-movers that have been paid-off to buy a new property

We identify movers through either "chains" or "pairs" of transactions. "Chains" of transactions are illustrated in Figure 12. We define "chains" as a set of three transactions: first, borrower a taking out an initial loan 1 to move into property x; second, the same borrower a (i.e. the home-mover) taking out a new loan 2 when he moves out of property x to property y; and third, borrower b taking out loan 3 to move into property x within a few days after borrower a leaves x. To identify a mover chain we use a two-step matching algorithm. The first one to match the mover's initial loan 1 to the new loan 2 using the borrower a's date of birth (see Figure 12). And a second algorithm to match loan 3 from borrower b purchasing the mover's old property x to the mover's old loan 1 using the property postcode. All the possible matches are represented by connecting lines in the diagram which we refer to as "edges".

[18] Developing a probability of default (PD) model is not the main objective of this work. Therefore, we use a simplified approach as defaults are generally a very small percentage of the total loans in the stock. However, our model is flexible enough to utilise a more sophisticated PD model.

[19] A mortgage lender has much more discretion on when to record a default, but it should generally not be later than six months after any repossession. See `https://debtcamel.co.uk/debt-default-date/`

Figure 12: Illustrative mover chains



Identifying "chains" allows us to track movers with some certainty, but this method does not work if a link in the chain is missing. This can be the case, for example, if there is a last-time seller or cash-buyer among the counter-parties, as their transactions are not be captured in the flow data (similarly if the property is bought with buy-to-let mortgage). Therefore, we also identify movers by matching "pairs" of mortgages using the borrower's date of birth only, ignoring the possibility of a new buyer for the same property. We apply this second relaxed matching technique only for those home-movers we could not find previous mortgage. At every stage of these matching techniques we find multiple possible candidates for each home-mover mortgage. To pick the best candidates in a way that is globally optimal, we design and implement a weighted bi-partite graph-matching algorithm (Cormen et al. (2009)) that maximises a global objective function based on scores capturing the quality of the matchings.

For each edge, we allocate a score signifying how closely a corresponding candidate matches a home-mover. The scores are calculated based on the difference in equity, house price, geographical distance etc. corresponding to the two loan entries. For example, we expect the majority of movers to find a newer home closer to their old

home. A score function captures these attributes and includes a constraint on the length of tenure as it is less likely that borrowers move homes very soon after acquiring a property. To address multiple matches and yet find the best list of candidates, we used a push-relabel [Goldberg and Tarjan (1986)] algorithm.

6.4 Identifying loans which have been re-mortgaged

We match re-mortgagors' latest loans with their previous loans using postcodes and borrowers' date of births. Where there are multiple matches, the best candidate is picked based on qualitative characteristics, including change in income and property characteristics (such as number of bedrooms, type of house or whether there is a garage).

A large proportion of mortgages are owned jointly by more than one borrower, however the flow data only capture the main borrower's date of birth. This means that when someone re-mortgages, there is a possibility that the main and secondary borrowers are switched, such that the date of birth between the original loan and re-mortgage changes. To circumvent this issue, we match re-mortgages with previous mortgages using only their postcodes (only for re-mortgages that we could not track by the above approach). We also impose conditions to reduce the size of possible candidates: the age difference between main and secondary borrowers is arbitrarily restricted to at most 10 years, the property value should not increase by more than 10% per year and decrease by more than 5% per year, the owner is staying in the property for at least one year before re-mortgaging, and if the dwelling type is known, it should be the same as the property is expected to be the same. Once the pairs of candidates are restricted to satisfy all the above mentioned criteria, we use an unweighted bipartite matching technique to find the best possible one to one map (similar to the approach we have used to track home-movers).

6.5 Borrower income projection

Using two household-level panel surveys, *British Household Panel Survey* (2004-2009) and *Understanding Society* (2009-2015), we categorise mortgagors into 7 mutually exclusive groups by age, employment status and income level:

1. Age 20-65 & Employed & Income £0-20k
2. Age 20-65 & Employed & Income £20-40k
3. Age 20-65 & Employed & Income £40-80k
4. Age 20-65 & Employed & Income £80k+
5. Age 20-65 & Self-employed & Income £0-50k
6. Age 20-65 & Self-employed & Income £50k+

7. Age 65+ and/or retired

We choose these groups as they have different (log) annual income change distributions. Kolmogorov-Smirnov tests [Daniel (1990)] also confirm that these distributions are significantly different. We fit normal distribution on log income change for each group for each consecutive year pair.

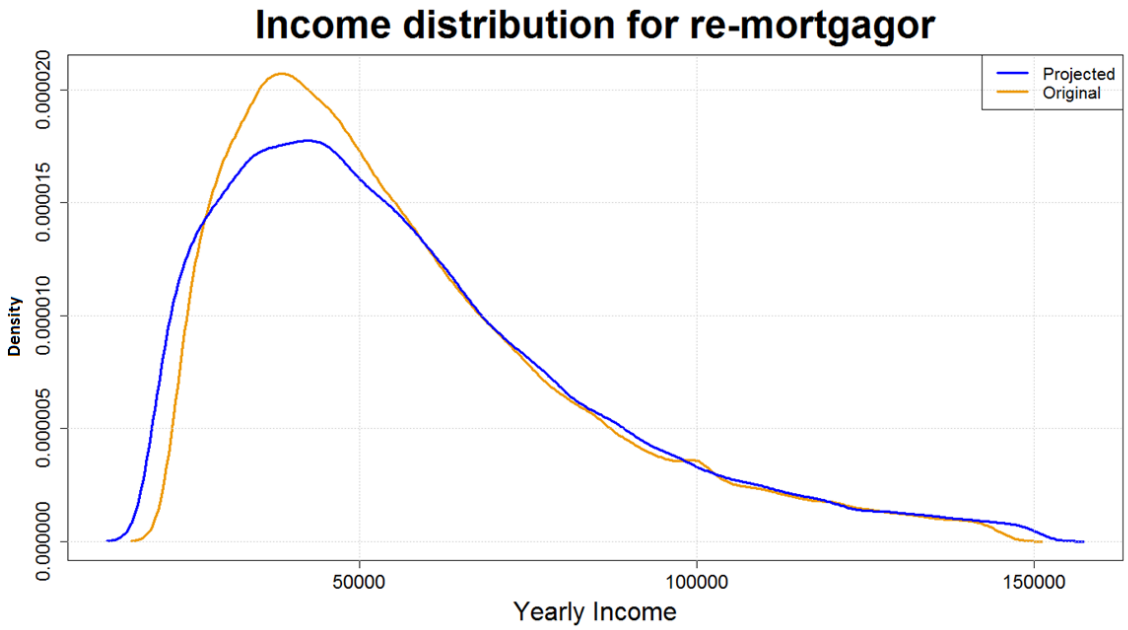
Similarly, we group our stock estimation into the same 7 groups. We adjust origination income for self-certified loans which consitute around 40-50 percent of loans originated before the financial crisis. There is evidence [BBC (2004)] that borrowers often lie when their incomes are not verified specially when they are self-employed. Hence, we assume that self-certified borrowers inflated their income by 25 % at origination when they are self-employed and 20 % when they are employed. Then for groups 1-6 we project origination income annually by applying the distribution produced for the corresponding group and year from surveys.

Our approach allows mortgagors to switch from one group to another group over time. For example, if in a given year a negative income shock leads a borrower moving to a lower income group, then for the next year this borrower is assigned to a random income change from the distribution for the lower income group in the respective year.

Similarly, some borrowers can move into 65+ age and/or retired bucket over time. If we do not take into account income falls due to retirement and continue growing their income after they move into this bucket (inline with the income change distribution for group 7), we may get significantly higher income estimation. Unfortunately, we do not have information regarding how an individual's income reduces after retirement compared to his/her pre-retirement income. Therefore, we keep their income constant at origination. Given that we have a very small number of observations in this category in our estimation, this would not affect our overall projected income distributions.

We validate our approach using data on the re-mortgagors identified in the first stage of our estimation. For re-mortgagors we observe incomes at two points in time: when the initial loan was taken out and when the borrower has re-mortgaged. For these borrowers we project forward their income from origination and compare it against their actual income reported when they re-mortgaged. The income distributions of our projection and observed incomes of re-mortgagors are shown in Figure 13. As shown by the figure, our projection has a slightly longer left tail. This means we marginally underestimate the income of the borrowers at the point when they re-mortgage. This is not surprising if borrowers who are financially better off than average borrowers find it more easy to get a new re-mortgage deal and are therefore more likely to re-mortgage. Our approach does not consider specific factors related to re-mortgaging in our income projection model (e.g. different income distributions for re-mortgagors).

Figure 13: Income Projection



6.6 Outstanding loan amount projection

The calculation of outstanding loan amount at any point in time from origination loan value is based on several assumptions. We use information on mortgage interest-rate, interest-rate deal (fixed, flexible, tracker and discount), deal end date, repayment type (capital and interest, interest-only and mixed), and mortgage term at origination. Below explanations are for 2015 stock estimation but we apply the same approach when estimating the stock for other years.

In the flow data, for mortgages with fixed, tracker and discount deals we have information on their deal end dates. However, this information is missing for significant amount of these mortgages, so we first start with imputing deal end dates for these loans.[20]

We assume that fixed, tracker and discount mortgages switch to flexible rate when their terms end. According to interest-rate deal types and the end date of these deals (whether deals end before or after 2015), mortgages are grouped into five categorizes:

1. Flexible rate mortgages since origination

2. Fixed rate mortgages since origination (deal end date is after 2015)

[20]If deal end date is missing, we use MoneyFacts data. It provides information on deal end dates for mortgage products. See `https://moneyfacts.co.uk/`. If it is still missing, then we use some aggregated statistics from the flow data to impute missing deal end date.

3. Tracker and discount mortgages since origination (deal end date is after 2015)

4. Fixed rate mortgages at origination but switch to flexible rate when their terms end (deal end date is before 2015)

5. Tracker and discount mortgages at origination but switch to flexible rate when their terms end (deal end date is before 2015)

For computational simplicity, we calculate average monthly payments during relevant periods and compute outstanding amounts as of 2015. Alternative could be calculating payments for each month and updating outstanding amounts monthly between origination to 2015, however this would be computationally time consuming. For the group of mortgages (1-3), average monthly payments are calculated for the entire period from origination to 2015. For the group of mortgages (4-5), two different average monthly payments are calculated for the periods: (i) between origination and the end of the deal term, (ii) between the end of the deal term and 2015.

To calculate average monthly payments, we need to adjust the interest-rate information due to two reasons: (i) interest rate is subject to change over the course of the mortgage for flexible, tracker and discount rate deals, and (ii) we assume when deal term ends fixed, tracker and discount rate mortgages switch to flexible rate. The adjustment is done according to interest-rate types:

- Flexible rate mortgages: interest rates are adjusted using Standard Variable Rates by mortgage lenders between 2005-2015 reported to Bank of England.

- Fixed rate mortgages: during their deal period, monthly payment calculations are based on fixed interest rate at origination. If their deal terms end before 2015, then average flexible interest rate is used for monthly payment calculations between the end of the deal and 2015.

- Tracker and discount loans: similar to fixed rate mortgages, if their deal term ends before 2015, then average flexible interest rates are used for monthly payment calculations between the end of the deal and 2015. However, during the deal period, we also need to adjust the tracker and discount interest rates as the underlying interest rates are subject to change. We calculate the spread at origination, which is the difference between origination interest rate and Bank rate at the origination quarter. We calculate the discount at origination, which is the difference between origination interest rate and Standard Variable Rates at the origination quarter. We assume that spread and discount at origination is fixed over the deal term. For tracker, we add spread to the average Bank rate over the period of consideration. For discount, we add discount to the average SVR over the period of consideration.

6.7 Property Value Projection

For property value calculations, we use the Land Registry's UK wide house price index, regional house price indices and Price Paid dataset. From the regional house price indices, we calculate house price change for all possible year-quarter pairs for all regions (i.e. London, South East England, Scotland etc.). From the Price Paid dataset, we repeat the same process for more granular geographical units such as county and NUTs [Nomenclature of Territorial Units for Statistics European Commission (2017)] by using a mapping from postcode to county and NUTs.

Origination property values in our estimation are then projected by using the median house price change of three geographical classifications: regions, county and NUTS. If the propert value is missing for a loan (suppose, the postcode is missing or invalid), the property value is grown according to the Land Registry's UK wide house price index.

Alternative to using the median of region, county and NUTs level house price changes, we use also the following hierarchical approach:

1. first use the NUTs level house price change (as that is the most granular geographical segregation),

2. if that is not possible, use county level house price growth (we have missing information for some NUTs but we have county level information available corresponding to some of them and that is the next granular level geographical area in our data),

3. if that is not possible, use region level house price growth,

4. if that is not possible, use UK wide house price growth (i.e. the postcode is missing or invalid).

Overall, the median looks slightly closer to the aggregate when we compare that with the WAS in the validation.

There is also a reason for using a simple median. Although from NUTs and county level house price changes we can capture more granular effects, these indices are created based on reported sale prices in the Price Paid dataset. Properties are generally sold when their prices are sufficiently appreciated, otherwise the owners wait for a better time to sell. Hence, these house price changes might be biased upwards. Moreover, smaller the geographic region is, the fewer number of properties will be sold per quarter. As we do not put any restrictions on property size or other factors, a fewer number of sold properties in a smaller geographical area will make the price movement more volatile. On the other hand, the regional indices are officially published and adjusted for some of these biases. That is why we do not give any extra priority to indices based on more granular geographical regions.

6.8 Validation: FCA stock data and Household Surveys

6.8.1 FCA stock data

We use the FCA stock data for our sample validation. However, we can not compare the FCA stock data with our estimation immediately. First we drop loans originated pre-2005 to make it comparable to the FCA flow data available. Then we match the data set with the flow data (PSD 001) as this will help us to get the LTV and LTI distributions at the origination for PSD 007 loans. We use that for the validation of our estimated stock. As there is no identifier to match the stock and flow data, we use a step by step approach to match these two. The matching criteria are:

Level 1 matching: Best quality match

First, we match the two datasets only when the following variables are exactly the same:

- Date of origination of the loan
- Date of birth of the borrower
- Loan value at origination
- Postcode of the property
- Lender Group[21]

Level 2 matching

Next, we relax some of the matching criteria for those unmatched loans as mentioned below:

- Opening dates or loan origination dates are at most one month apart
- Dates of birth of the borrower are the same
- Loan values at the origination are within 5% of what is reported in the stock
- Postcodes of the properties are the same
- Lender Groups are the same

Level 3 matching

For the rest of them, we relax the matching criteria further. Next, we consider only the following variables to be the same:

- Date of origination of the loan
- Date of birth of the borrower

[21]Rather than the exact bank id, we have used the bank group to get a better matching as loans can be switched within a group.

- Postcode of the property
- Lender Group

Duplicate matchings are handled by using a bipartite graph matching algorithm, as the one we use for the home-mover matching. But in this case there are not too many duplicates, unlike home-mover matching. We match around 90% of the FCA stock data with the FCA flow data using our approach.

Table 6 presents some summary statistics for the matched FCA stock data.

Table 6: PSD 007 statistics

Characteristics	Variables	Values from in FCA stock 2015 H1
Loan	Interest only share	18%
	Current average interest rate	3.09%
	Fixed ratetype share SVR ratetype share	42% 30%
	Average remaining Mortgage term	16 years
Property	London and South East location	32%
Borrower	Average age	45 years

6.8.2 Household Surveys

Understanding Society provides longitudinal data about subjects such as health, work, education, income, family and social life, and to understand the long term effects of social and economic change and of policy interventions designed to impact the well-being of the UK population. Understanding Society has been running since 2009 and has a sample size of around 40,000 households and 100,000 individuals. Each wave of the survey is collected over a period of two years in England, Scotland, Wales and Northern Ireland. The predecessor to Understanding Society was the British Household Panel survey (BHPS), which began in 1991. The final wave of the BHPS, wave 18, was collected in 2008. BHPS sample members have been incorporated into the Understanding Society survey from wave 2 onwards. There are some small methodological differences between the two surveys, but for the purpose of this work we marge them together.

The Wealth and Assets Survey (WAS) is a longitudinal survey of households and individuals across Great Britain. The primary purpose of the WAS is to fill the gap in the data by gathering information on four main areas of households' finances: property, physical, financial and pension wealth. Additionally, the WAS contains information on labour market conditions, business activities, sources of income, attitudes

towards spending and borrowing, and financial distress including debt and arrears on a variety of credit agreements and house bills. At present, there are four waves are available (2006-8, 2008-10, 2010-12 and 2012-14). Wave 4 has a sample size of around 20,000 households and 40,000 individuals.

The Bank of England / NMG Survey (NMG) is carried out by NMG Consulting on behalf of the Bank of England. The survey covers households in Great Britain. It contains information on mortgage and unsecured debt, income, financial and housing wealth, spending and saving. Also in each survey ad-hoc questions on topics of current policy interest from the Monetary Policy and Financial Policy Committees of the Bank of England are included. The NMG survey has been run on an annual basis since 2004. Since 2014 it has been run biannually, in April and September of each year. The survey was initially carried out face-to-face, covering about 2000 households in each wave. Since 2012 the survey has been run online; the lower cost of the online survey has allowed the sample size to be tripled to around 6000 households.