

Residential Mortgage Prepayments: Issues, Model Structure and Implementation

Quantitative Finance, Atlanta

5 Mar 2001

Abstract

Following a brief compilation of internal, Street, and vendor ideas on prepayment modeling, the current prototypes for QF's fixed rate, ARM, and hybrid prepayment models are described. Changes are inevitable.

Part I

A Qualitative Survey of Modeling Issues

Part one of this paper seeks to distill and summarize some of the better (and often anecdotal) ideas generated internally and gleaned from published papers. In particular, the content is directed toward traditional GNMA, conventional and jumbo mortgages, although many of the concepts extend naturally into subprime, mobile home, and other markets.

Residential mortgage prepayments are often carved into four parts: housing turnover, refinancing, defaults, and curtailments, with the latter two generally receiving little attention. This pattern will persist herein.

1 Housing Turnover

The turnover component of prepayments captures primarily non-economic and demographic effects: homeowners prepay mortgages due to job relocation, housing upgrades, divorce, death, etc. While some of these behaviors are not independent of prevailing economic conditions, e.g. housing upgrade costs rise with interest rates, these drivers imply a base level of prepayments prior to considering refinancing.

1.1 Seasonal Effects

A seasonal element to housing turnover, generally attributed to weather, holiday, and school calendar effects, is often modeled such that a seasonal minimum occurs in January and a maximum occurs in June. Such a pattern might be estimated from home sales data or taken from published sources¹.

¹The National Association of Realtors and some Wall Street firms have published seasonal factors.

1.2 Expected Tenure

This component of a housing turnover model is often addressed with a PSA-like seasoning ramp, suggesting a linear progression toward a stable annual turnover rate². The relative length of the ramp as well as the level of this stable turnover rate are indicative of differences in expected tenure among loan types. Considerations for modeling expected tenure include:

- Loan type: ARM and balloon loans tend to be favored by “fast movers.” FHA/VA loans are made to a demographic group which empirically exhibits less mobility. In contrast, 15 year loans may be favored by a more sophisticated and mobile population with a better credit rating.
- Type of property: single family structures are indicative of longer tenure than condominiums, etc.
- Points paid: large rate buydowns do not make sense unless the homeowner plans stay in the mortgage for an extended period of time. Such a points effect is usually allowed to decay over time.
- Loan purpose: if a loan is the product of refinancing, the loan can be thought of as *preseasoned*. This implies that the expected tenure in the home may be shorter than in the case of a purchase. Furthermore, note that if the curve was relatively flat at the time of the refinancing, “fast movers” that would normally apply for balloon or hybrid ARM loans may instead have taken 30 year loans. Such a loan, therefore, would have a faster turnover profile than one originated in a steeper curve environment.
- Prepayment penalties accepted: as with points paid, acceptance of prepayment penalties suggests a propensity to remain in the mortgaged property for the term of the penalty provision. It is possible that a spike in prepayments will be observed immediately following the end of the penalty period as pent up turnover is released.
- Loan size: large loans may imply reduced tradeup potential and longer time to market, whereas small loans may suggest the opposite effects. However, corresponding borrower demographics (mobility, net worth, etc.) make such contentions less than certain, as jumbo whole loans are on average thought to exhibit greater turnover than conventionals.
- Geographic location and mobility: California is often said to have a more mobile demographic profile. An alternate viewpoint attributes faster speeds in California to other variables, such as larger loan size and popularity of ARM products, yet this does not seem to be supported by the internal California data. Pre-merger BankAmerica models also captured geographic differences between northern and southern California.
- Geographic location and LTV: Regional housing market differences clearly have an impact on turnover behavior.
- Geographic location and transactions costs: Applicable at the turnover and refinancing levels, the difference in taxes and fees across geographic entities is said to have a significant impact on relative prepayment behavior.

²A strict linear ramp to a constant level may prove too simplistic, as empirical data may suggest that the post-ramp turnover rate should decay over time. Self-selection effects may be present as homes in higher growth areas leave the pool.

- Limited equity: High LTV loans and loans which were equity take-out refinancings are more likely to reach low or negative equity conditions in housing market slumps. Borrowers in such situations may anticipate longer tenures and associated home price appreciation as a source of equity.
- Other: Relocation loans may exhibit initial stability followed by an acceleration in prepays. Low documentation borrowers may be self-employed with strong geographic ties, and persistent aversion to full documentation may limit access to new credit as well as discourage new home purchases. Investor properties may turn over more rapidly than owner-occupied properties, as investors may be more likely to take gains resulting from housing market appreciation.

1.3 Economic (Dis)incentives

While economic incentives are commonly discussed in the context of refinancing-related prepayments, housing turnover is clearly affected by similar considerations.

- Prevailing mortgage rates: housing turnover is dampened in environments where the purchase of a new home will involve a mortgage with a higher interest rate: this is often referred to as a "lock-in effect." The effect of this disincentive may fade somewhat over time, however, as many non-economic reasons for moving cannot be indefinitely postponed. Conversely, refinancing opportunities will likely amplify housing turnover.
- Loan size: as the prevailing rate disincentive is a linear function of size, jumbo loan prepayments would be expected to dampen relatively more in such scenarios. This is a potential explanation for the anecdotal claim that jumbo loans have prepayment functions which are *steeper* than conventionals.

1.4 Broad Economic Effects

The state of the economy, trends in housing values, and consumer confidence surely matter, but will likely defy attempts at accurate forecasting or meaningful incorporation into a stochastic interest rate model of low dimension.

As of early 2000, comparisons of published prepayment projections across multiple models suggest substantial differences in opinion over "out-of-sample" speeds in rising rate scenarios. Such differences are driven by varying assumptions of housing market performance in a high mortgage rate environment rarely seen in the past decade.

2 Refinancing

The portion of prepays which would seem to lend itself to a simple and elegant economic interpretation is in reality complex and challenging, especially for ARM's. Topics to be considered when addressing refinancing include:

- Prevailing mortgage rates:

- ARM vs FRM: While most sophisticated mortgage shops will express a certain comfort level regarding fixed rate prepayment models, practitioners are often less sanguine regarding ARM models. The heterogeneous nature of production (relative to agency fixed rate mortgages), implicit borrower expectations of the direction of interest rates, borrower risk aversion, and refinancing alternatives along the entire yield curve (ARM-to-ARM and ARM-to-FRM refinancings) are among the sources of prepayment modeling difficulties³. Note that borrower decisions to refinance an ARM loan will incorporate features such as caps, relative margins and teasers: these should be incorporated into the prepayment model.
- Refinancing alternatives: With 15 or 30 year fixed rate loans, a current coupon rate at a comparable point along the yield curve is likely to capture most of the refinancing incentive. In steeper yield curve environments, however, refinancing may accelerate more than in parallel scenarios: sophisticated fixed rate borrowers may roll down the curve to capture greater savings. With ARM loans (and likely balloons), curve behavior is even more important. Models will need to consider refinancing into new ARM's as well as fixed rate loans. The ARM-to-FRM refinancing component is also a function of the absolute level of interest rates, where borrowers may forego greater ARM-to-ARM savings in favor of a fixed rate near recent historical lows.
- Lags: When forecasting refinancing behavior, it is generally accepted that a lag in rates of one to two months is needed to capture the length of the refinancing process. The lagged interest rates should be some moving average to capture intramonth rate movements, with large dips receiving potentially larger weights as prospective borrowers lock in loan rates.
- Coupon ratio or difference: Measurement of refinancing incentive may be considered as a function of the absolute difference between the loan rate and prevailing mortgage rates or as a ratio of the two. The latter approach has been proposed to address the notion that the economic incentive to refinance is not only a function of the rate difference but the level of rates as well⁴. The marginal explanatory power of the ratio variable instead of a rate difference is an empirical question. Note that the ratio approach implies an exploding refinancing incentive as rates trend toward empirical lows, as a 2 percent rate difference in a 4 percent environment ($6/4 = 1.50$) could imply sharply higher prepayments than in a 6 percent environment ($8/6 = 1.25$): while it is theoretically appealing to include discounting in the borrower's decision process, this relative incentive may be too strong in low rate environments, particularly at levels which were not observed in the empirical sample.
- ARM-to-ARM prepayments: ARM borrowers need not refinance to capture downward movements in the short end of the curve, as such resets occur naturally within the existing loan. Hence, ARM-to-ARM refinancing must be explained by factors such as teaser capture

³Westhoff and Srinivasan claim that the slope of the curve is important only to marginal loans which are only refinaneable with a roll-down into ARM or balloon products. (Bear Stearns, *The Next Generation of Non-Agency Mortgage Valuation Models*, March 1999.)

⁴This point is often tossed about rather loosely. In fact, a given coupon differential implies greater economic savings at higher rate levels prior to discounting. Whether discounting effects will overcome this payment savings effect depends upon the assumption made regarding the borrower's decision horizon.

and equity takeout.⁵ A pronounced prepayment spike is commonly observed around rate adjustment dates in certain ARM sectors, particularly around the initial reset. This effect has been attributed to some combination of teaser effects, short tenure selection bias, and equity takeout plays.

- Tiers of borrowers:
 - Points and margins: Borrowers who buy down loan rates via payment of points may have better credit characteristics, lower LTV ratios, and sufficient financial savvy to refinance aggressively when the opportunity arises. While data are not always readily available on points paid, a common proxy considers the loan rate versus prevailing mortgage rates at origination. A closely related effect involves borrowers with higher-than-average margins at origination, as these borrowers may be less responsive to refinancing opportunities (due to poorer credit, higher LTV ratios, etc.) Models which incorporate this effect generally allow it to decay over time.
 - Loan size: as the dollar savings from refinancing is a linear function of loan size, loans should exhibit relative responsiveness consistent with balance differences. Some of the observed differences in refinancing responsiveness of FHA/VA, conventional, and jumbo loans will undoubtedly be explained by loan size differences. On a relative basis, refinancing transactions costs are lower for larger loans.
 - Expected tenure again: As the benefits of refinancing involve weighing the present value of payment reductions against transactions costs, properties which exhibit rapid housing turnover (such as condominiums) will realize smaller savings from refinancing and hence tend to respond more slowly to refinancing opportunities. *Preseasoning* may also be a consideration if a pool has a substantial percentage of loans with a refinancing purpose.
 - Penalties: Prepayment penalties should be explicitly modeled as a cost or *hurdle* to be considered in a refinancing decision. It is possible, however, that borrowers who accept such provisions will be less responsive in general, prepaying more slowly than the economics of the penalty provisions would suggest.
 - Equity takeout opportunities: leveraged borrowers who have experienced high levels of home appreciation may exhibit a tendency to respond quickly to refinancing opportunities.
 - Origination channel: banking center, mortgage company, and correspondent originations may exhibit varying degrees of borrower financial savvy as well as varying exposure to proactive refinancing offers.
- Burnout: Prepayment models usually include a component aimed at capturing the tendency of successive refinancing waves to dampen through time, as responsive borrowers leave a pool of loans and the remaining borrowers are less willing or able to follow. A number of approaches are possible:

⁵Downward resets in rate caps are a potential byproduct of ARM-to-ARM refinancing, but borrowers are unlikely to view the caps as a primary motivator.

- Continuum of responsiveness: Perhaps the simplest approach is to assume that the borrowers in a pool of mortgages have a refinancing response function which is dampened as the ratio of remaining borrowers to original borrowers (the *Survival Ratio*) goes down. For example, after 25% of the loans have refinanced, the remaining 75% may prepay at 90% of the previous level. When the survival ratio drops to 50%, the refinancing response might dampen to 75% of the original level. When the survival ratio reaches extremely low levels, refinancing responses dwindle even further.
 - Discrete tiers of responsiveness: A fairly common approach explicitly models tiers of borrowers, potentially including segments which cannot respond due to credit or LTV issues, segments which respond with minimal incentive, and some segments in between these extremes. Each segment may have its own refinancing response function⁶. Some models incorporate dynamic movements between segments, although this suggests excessive complexity.
 - Burnout as an illusion: At least one Wall Street quant⁷ has suggested that burnout is almost completely the result of loan level characteristics such as those listed above: points paid, margins, size, LTV, documentation level, etc. Natural segmentation into subpools, combined with a comprehensive model addressing these characteristics, is claimed to explain the burnout effect.
- Media effects: Some prepayment analysts point to *media effects*, suggesting that as rates near recent historical lows, refinancing responsiveness is amplified for a short period of time. This feature may be incorporated to "heal" burned out loans. A "look-back function" has been suggested⁸ as an implementation mechanism, driving refinancing behavior as a function of rate level and time since the last similar opportunity⁹.
 - Costs and convenience: Certain loans are more readily refinanced: convertible ARM loans, for example, may prepay more rapidly than similar nonconvertible loans *IF* a refinancing opportunity exists and the loan is within its convertible window (usually years 1 through 5.) Some branch-originated loans include rate modification provisions: these allow a borrower to reset the loan rate to market under certain conditions for a fee. Mortgage banking industry effects are relevant as well, as low cost refinancing programs accompanied aggressive refinancing campaigns in the early 1990's.
 - Loan assumptions: FHA/VA loans are assumable, allowing a purchaser of a home to assume the mortgage of the previous owner. This will only occur, however, if the existing rate is favorable and the home has experienced minimal appreciation.

⁶Larger balance and higher coupon loans would naturally exit first, leaving a pool of less responsive loans.

⁷Steven W. Abrahams, *The New View in Prepayments*, Morgan Stanley, February 1997.

⁸Westhoff and Srinivasan, 1999.

⁹Higher refinancing responses are claimed to occur when a longer period of time has passed since a similar opportunity. This effect need not occur as a function of media coverage of historical troughs: equity takeout opportunities are likely to be greater with longer intervals between opportunities.

3 Defaults

While default modeling is an important topic in credit-sensitive contexts, the value added by addressing defaults as a separate component of prepayments is probably low relative to the development of solid turnover and refinancing models. When embedded into a prepayment model, default is often modeled as a simple function of age, LTV, and a few other credit variables. It should be noted that ARM loans are generally considered more likely to experience credit performance problems due to the potential for payment shocks and usage as an avenue to qualify for larger loans.

4 Curtailments

Partial prepayments are generally assumed to be immaterial, conveniently collapsing into the parameters of the turnover portion of a prepayment model. A possible exception, however, is the case of extremely seasoned 30 year loans, where the interest portion of a payment is relatively small (minimal tax advantages) and the borrower has likely accumulated personal wealth over the term of the loan.

Part II

Model Structure and Implementation

5 Overview

The approach taken by Quantitative Finance has traditionally involved the construction of a prepayment function similar to those commonly used by traders and dealers, with the associated parameters fitted to some "Street Consensus," imputed from market prices, or driven by assumption.

5.1 Statistical Approaches: Then and Now

In the past, econometric techniques were avoided for a number of reasons:

- the data and resources required were not readily available without substantial costs;
- the nature of the FRM portfolios often analyzed was assumed to be similar to TBA pools, allowing inexpensive leveraging off of published forecasts;
- the preferred functional forms exhibit non-linearities and interactions between explanatory variables that are non-trivial to address with standard econometric machinery, and
- finally, in light of the previous points, skepticism existed as to the marginal value of an econometric model over an implied model.

It is clear, however, that the use of implied forecasts is best limited to homogeneous, liquid product sectors where market participant have adequate data and incentives to estimate and publish prepayment analysis. As a significant portion of the Bank of America mortgage portfolio falls outside of such sectors

of the market and is therefore in need of specialized models, it seems appropriate to attempt to integrate the empirical and prospective approaches.

Recently data issues have been addressed post-merger with a concerted effort to collect and organize data from both the East and the West bank. As a result, data sets for both fixed and adjustable rate mortgage pool histories have been created. In addition, data for fixed rate mortgages from outside sources (notably, EJV) has become available as well. Quantitative Finance is currently building a purely empirical model in addition to using statistical methods to evaluate the performance of production models.

5.2 Lessons from Derelict Models

The past decade of prepayment research would seem to teach a number of lessons.

- First, the structure of the mortgage market is dynamic: care must be taken to account for structural changes over time;
- Second, prepayments are driven by numerous variables other than the general level of interest rates: the strength of the economy, consumer confidence and leverage, housing prices and prevailing LTV levels, debt consolidation opportunities, the slope of the mortgage curve, etc. As these drivers are often missing from deterministic prepayment models due to difficulty in forecasting the values of such variables, it is necessary to make subjective judgements as to whether a model should fit empirical behavior over long periods where such variables are changing.
- Finally, given such problems in building a model with econometric techniques, the ultimate product of such efforts should be sufficiently transparent to allow careful inspection and testing of the sensitivity of the performance of the model to changes in "unmodeled" variables.

The usual components discussed in prepayment papers are incorporated in the current model specification: seasoning (slower prepays early in the life of a loan), seasonal variation (higher prepays in warmer months,) refinancing behavior, and burnout (refinancing responses dampen as opportunities are missed.) The manner in which these components are incorporated will vary as a function of the type of mortgage considered (ARM, hybrid, fixed, balloon, etc.)

6 Key Concepts

Before wading into the equations, a discussion of some notable issues may prove to be helpful.

6.1 Functional Form

The model employed includes both additive and multiplicative elements. The additivity is aimed at segmenting behavioral effects such that refinancing opportunities do not face a seasoning ramp or seasonal variation and that non-refinancing prepayments¹⁰ (deaths, divorces, relocations, upgrades, etc.) do not burn out in the same fashion as refinancing. Early prepayment models allowed these effects to blend.

¹⁰These prepay drivers are collectively referred to herein as the *demographic* element of prepayments.

6.2 Forecasting Mortgage Rates

Analysis of mortgages is commonly performed in the context of some non-mortgage benchmark rate, such as CMT or CMS.¹¹ As prepayment forecasts are generally driven via benchmark mortgage rates, some assumption must be made concerning the dynamics of mortgage rates as a function of the non-mortgage benchmark, e.g., the 10 year swap rate.

Rather than build a simple regression model of the mortgage spread, Quantitative Finance makes the following assumption:

$$S_t = S_{t-1} + \gamma(S_\mu - S_{t-1}),$$

where

- $S_t \equiv$ the time t spread between the mortgage rate and a non-mortgage index (10 year CMS).
- $S_\mu \equiv$ the long-term expected value of the spread.
- $\gamma \equiv$ the speed at which the spread returns to its long-term expected value.

Thus, the prospective mortgage basis behavior is easily understood and may be subjected to exogenous shocks via S_μ and γ . A simple econometric model of the spread may be pursued at some point, with care taken to maintain model stability and transparency.

7 Fixed Rate Mortgage Prepayments

Fixed rate prepayments are assumed to be separable into a turnover component and a refinancing component. In each component, the current rate environment enters into a borrower's decision via an economic savings calculation: $Savings_t(\cdot)$ is a function of rate differential ($G - (C_t + S_t)$), loan size, borrower evaluation horizon, and remaining term T .¹² Note that this quantity may be negative, indicating a *disincentive* to terminate a loan for non-financial reasons.¹³ Define

- $G \equiv$ gross mortgage loan rate.
- $T \equiv$ weighted average maturity in months.
- $C_t \equiv$ current reference rate (e.g. 10YCMS) at month t . (May include a lag.)
- $Savings_t(\cdot) \equiv$ approximation for the time t economic value of refinancing.

7.1 Demographic or Housing Turnover Prepayments

Three multiplicative elements are used to capture prepayment behavior in the absence of a refinancing incentive. Note that the demographic component is not independent of a rate (dis)incentive, as housing

¹¹It is increasingly unusual for Treasury rates to be used as benchmarks for mortgage rates or OAS calculations. The "decoupling" of Treasury and spread products in recent years has led to widespread use of CMS (Constant Maturity Swap) rates as the preferred benchmark.

¹²See the appendix for additional discussion.

¹³The prepayment literature often dubs this the "lock-in effect."

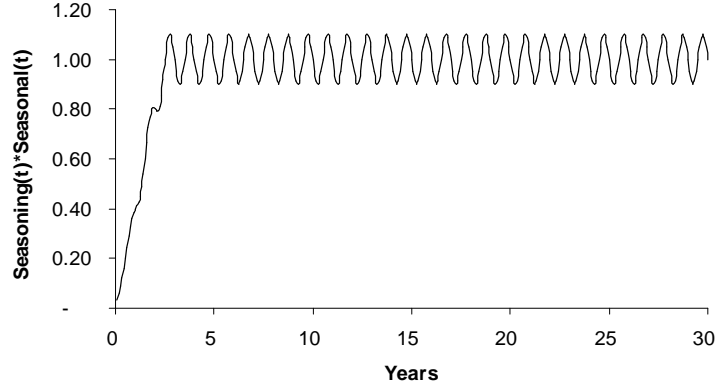


Figure 1: Seasoning and Seasonal Components

turnover is undoubtedly affected by the level of mortgage rates.

$$\begin{aligned}
 SMM_t^D &\equiv \text{demographic portion of SMM at time } t. \\
 Age_t &\equiv \text{months since origination.} \\
 MTS &\equiv \text{months until fully seasoned.} \\
 Seasonal(t) &\equiv \text{seasonal SMM multiplier as a function of calendar month.} \\
 \mathcal{F}^D(Savings_t(\cdot)) &\equiv \text{demographic SMM as a function of refi (dis)incentive .}
 \end{aligned}$$

The demographic component is given as:

$$SMM_t^D = Seasoning^D(t) * Seasonal(t) * \mathcal{F}^D(Savings_t(\cdot)).$$

A PSA-like seasoning ramp¹⁴, rising linearly from month 0 to month MTS, is given by

$$Seasoning^D(t) = \min\left(\frac{Age_t}{MTS}, 1.0\right) \mapsto [0, 1].$$

To capture a simple seasonal pattern, a look-up function may be used such that the SMM_t^D is scaled,

$$Seasonal(month(t)) \in \{0.70, 0.77, 1.05, 1.09, 1.14, 1.23, 1.11, 1.16, 0.99, 1.02, 0.91, 0.83\},$$

where $month(t) \in [1, 12]$ indexes a look-up table such as the set above.

Consider an "at-the-money" pool of mortgages where $C_t + S_t = G$.¹⁵ Assume that $\mathcal{F}^D(0.0) = 0.50$. The demographic prepayment function $\mathcal{F}^D(\cdot)$ suggests, therefore, that in the absence of an economic rate (dis)incentive, SMM_t^D will be 0.50 for a seasoned ($Age_t \geq MTS$) pool prior to seasonal considerations. If $Age_t < MTS$, then $\mathcal{F}^D(0.0)$ will be scaled downward by $\frac{Age_t}{MTS}$. Since $Seasonal(t)$ implies variation as a function of the current month of the year, $\mathcal{F}^D(0.0)$ will be scaled downward by as much as 0.70 in January and upward by as much as 1.23 in June.

¹⁴This shape assumes a constant after the ramp: some decay could be added. A multi-tiered shape may be used to capture the "slow mover" selection bias in prepayment penalty pools.

¹⁵For illustrative purposes, replace $\mathcal{F}^D(Savings_t(\cdot))$ with a function of pure rate differential $G - (C_t + S_t)$.

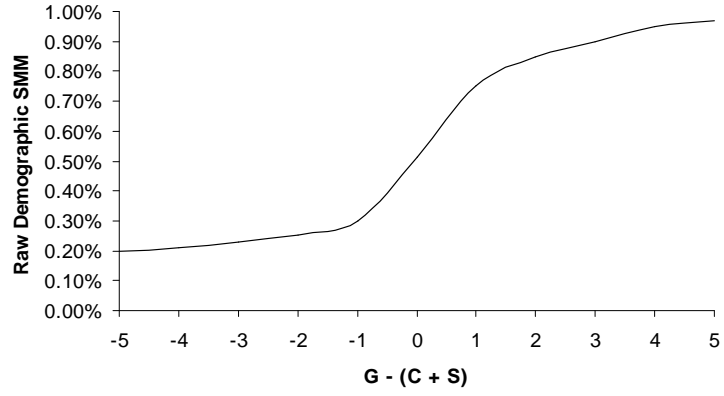


Figure 2: Demographic Prepayment Response $\mathcal{F}^D(G - (C_t + S_t))$

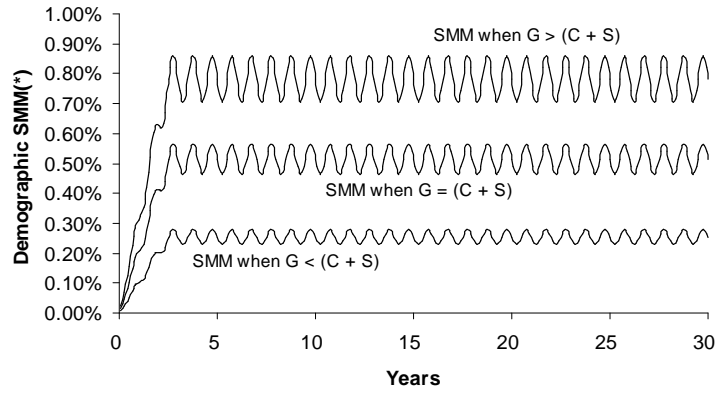


Figure 3: Demographic SMM at Several Rate Levels (SMM_t^D)

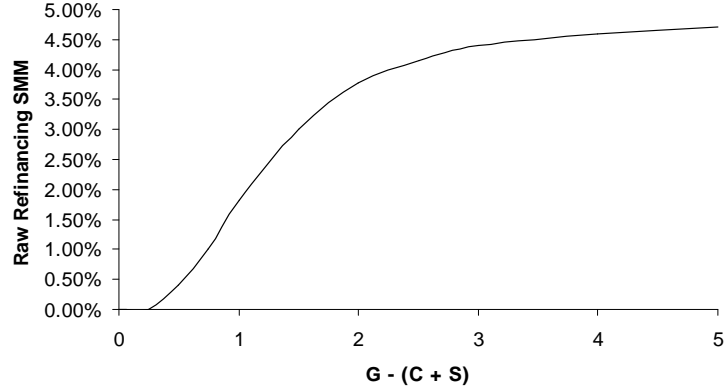


Figure 4: Refinancing SMM Response prior to Burnout $\mathcal{F}^R(\text{Savings}_t(\cdot), \text{hurdle})$

7.2 Refinancing-Driven Prepayments

Define:

$$\begin{aligned}
 SMM_t^R &\equiv \text{refinancing portion of SMM at time } t. \\
 \text{hurdle} &\equiv \text{refinancing cost hurdle.} \\
 \mathcal{F}^R(\text{Savings}_t(\cdot), \text{hurdle}) &\equiv \text{refinancing-driven prepay function.} \\
 \text{Survival}_t &\equiv \text{surviving loan proportion at month } t, \text{Survival}_t \in [0, 1]. \\
 \text{Survival}_t^\mu &\equiv \text{"expected" survival (after demographic runoff) at month } t, \text{Survival}_t^\mu \in [0, 1]. \\
 \mathcal{F}^B\left(\frac{\text{Survival}_t}{\text{Survival}_t^\mu}\right) &\equiv \text{refinancing burnout function } \mapsto [0, 1].
 \end{aligned}$$

The refinancing *hurdle* is used to introduce transactions costs and is a likely component of model enhancements aimed at analyzing pools which face prepayment penalties. The hurdle implementation assumes a fixed cost portion and a variable portion as a function of loan size.

The refinancing component is given as:

$$SMM_t^R = \mathcal{F}^B\left(\frac{\text{Survival}_t}{\text{Survival}_t^\mu}\right) * \text{Seasoning}^R(t) * \mathcal{F}^R(\text{Savings}_t(\cdot), \text{hurdle}).$$

$\mathcal{F}^R(\text{Savings}_t(\cdot), \text{hurdle})$ returns a refinancing-driven SMM level for a pool of mortgages which has seen no previous refinancing opportunities. Once refinancing opportunities are experienced, this component will be subject to burnout via $\mathcal{F}^B(\cdot)$. A separate seasoning mechanism $\text{Seasoning}^R(t)$ is included to add flexibility to the refinancing response functional specification.¹⁶

The burnout function dampens $\mathcal{F}^R(\text{Savings}_t(\cdot), \text{hurdle})$ as a function of observed refinancing behavior.¹⁷ Survival_t^μ is assigned the "expected" percentage of surviving loans at time t by assuming the

¹⁶For now assume $\text{Seasoning}^R(t) = 1.0$.

¹⁷Recent prepayment literature has emphasized the assessment of prepaes at the loan level as opposed to pool level: loan characteristics which lead to a distribution of refinancing responsiveness are modeled directly. Loans with sophisticated borrowers, large balances, low LTV/high equity takeout potential, and wholesale pedigrees are expected to respond quickly,

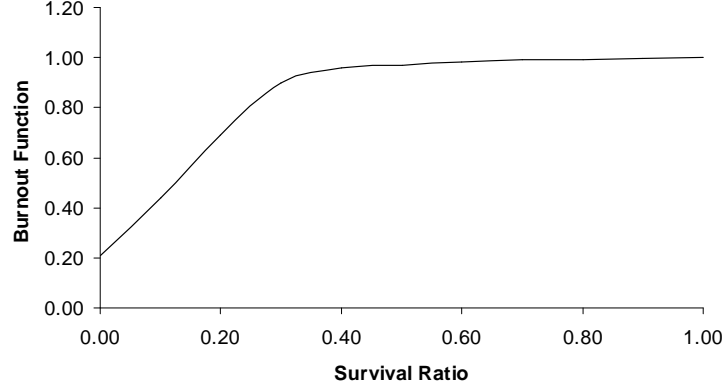


Figure 5: Refinancing Burnout Function $\mathcal{F}^B\left(\frac{Survival_t}{Survival_t^\mu}\right)$

pool has experienced SMM_t^D (as defined in the previous section) from origination to time t . To simplify the SMM_t^D calculation, it is assumed that $C_i + S_i = G, \forall (i < t)$. Thus,

$$Survival_t^\mu = \prod_{i=1}^t [1 - Seasoning^D(i) * Seasonal(i) * \mathcal{F}^D(0.0)] .$$

The actual percentage of surviving loans, $Survival_t$, is divided by $Survival_t^\mu$ to create a variable to explain burnout behavior.¹⁸

7.3 Demographic and Refinancing-Driven Prepayments

Combining the two components,

$$SMM_t = scale_t * (SMM_t^D + SMM_t^R),$$

where $scale_t$ is an exogenous shock facility and

$$SMM_t^D = Seasoning^D(t) * Seasonal(t) * \mathcal{F}^D(Savings_t(\cdot))$$

$$SMM_t^R = \mathcal{F}^B\left(\frac{Survival_t}{Survival_t^\mu}\right) * Seasoning^R(t) * \mathcal{F}^R(Savings_t(\cdot), hurdle)$$

The additivity accomplishes the aforementioned goal of modeling complex interactions correctly: refinancing opportunities are never seasonally-adjusted or dampened by a turnover-related seasoning ramp, and refinancing burnout does not effect the turnover behavior.

whereas loans with low balances, high LTV, credit impairment, documentation disincentives, or short borrower tenure would respond more slowly. The "traditional" burnout approach described in this paper is a proxy for such direct models of the distribution of loan characteristics. While the loan level approach has some appeal, the design and execution of such an approach comes at considerable expense in terms of data and computational requirements.

¹⁸Note that if burnout is treated purely as a function of $Survival_t$, omitting division by $Survival_t^\mu$, then an older, low-coupon pool of loans which has never seen a refinancing opportunity could experience a burnout effect based upon empirical demographic prepayment behavior. This possibility is precluded in the model via the survival ratio.

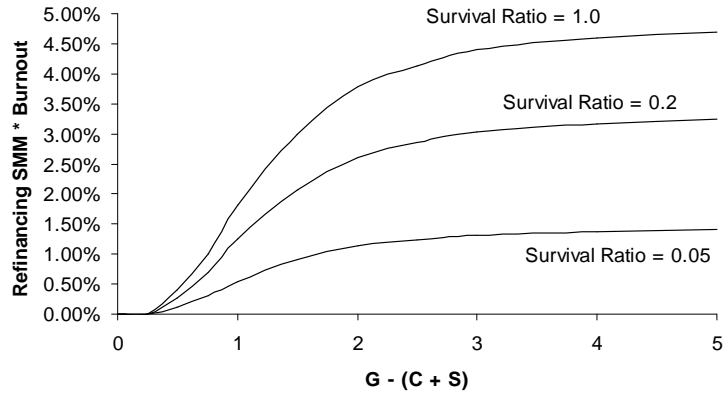


Figure 6: Refinancing Response with Burnout Adjustments

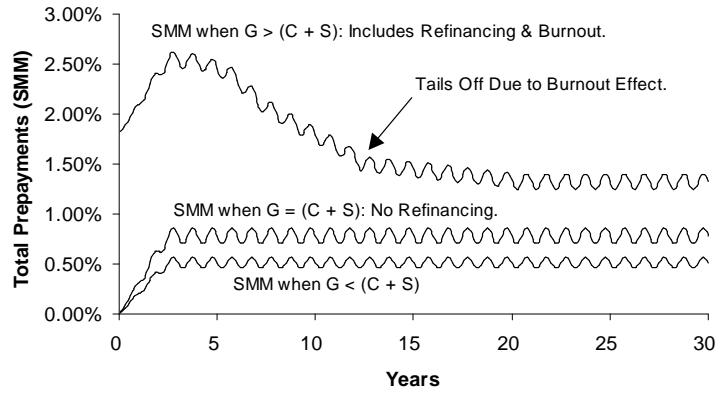


Figure 7: Combined Prepayment Effects

8 Adjustable Rate Mortgage Prepayments

ARM prepayments require numerous extensions to the preceding fixed rate model, but are addressed in a similar fashion. While the demographic functional form is the same as with fixed rate mortgages, the parameters reflect that ARM's generally season more quickly and have a higher and flatter demographic prepayment response function $\mathcal{F}^D(\cdot)$. For simplicity, the rate dependence of demographic prepaays is addressed via a FRM reference rate differential.

8.1 Refinancing-Driven Prepayments

Define:

- $G_t \equiv$ gross ARM current coupon.
- $C_0 \equiv$ original FRM reference interest rate, e.g., 10YCMS.
- $C_t \equiv$ current FRM reference interest rate.
- $S_0 \equiv$ original FRM reference spread.
- $S_t \equiv$ current FRM reference spread.
- $I_0 \equiv$ original ARM floating rate index.
- $I_t \equiv$ current ARM floating rate index.
- $E \equiv$ Borrower evaluation horizon in months.

The refinancing term has several components: an ARM-to-FRM component, a curve-steepening multiplier, and an "up-and-flatten" component.¹⁹

In falling rate environments, assume that ARM borrowers measure refinancing opportunities by comparing an original reference fixed mortgage rate ($C_0 + S_0$) to the current market fixed rate ($C_t + S_t$). Economic savings from refinancing are calculated as in the fixed rate model described previously. An SMM is returned via a function

$$\mathcal{F}_{\downarrow}^R(\text{Savings}_t(\cdot), \text{hurdle}).$$

To capture the potential for ARM-to-ARM refinancing, assume that this level-driven refinancing behavior is modified by a slope multiplier $\mathcal{F}^S(C_t - I_t, C_0 - I_0)$.

The "up-and-flatten" component is only relevant when the level of rates is such that $I_t > I_0$. ARM-to-ARM refinancing is assumed to be non-optimal, but the ARM-to-FRM question remains : is it the case that the average ARM coupon $\bar{G}_t(t, t + E)$ anticipated over time horizon E is high relative to a fixed rate alternative?

$\bar{G}_t(t, t + E)$ incorporates all relevant contractual features of the ARM such as margin, periodic caps and floors, life caps and floors, etc. The current evaluation horizon E for anticipated ARM rates is five years, with forward index resets ($I_t, i > t,$) assumed constant at the spot level I_t . Refinancing behavior is driven by the difference between the anticipated average ARM rate and the market fixed alternative:

$$\mathcal{F}_{\uparrow}^R(\text{Savings}_t(\bar{G}_t(t, t + E), (C_t + S_t), \cdot), \text{hurdle}).$$

¹⁹Note that the latter two components will not be fully exploited if used with single-factor valuation models.

8.2 Combining Demographic and Refinancing Components

Combining all components,

$$SMM_t = scale_t * (SMM_t^D + SMM_t^R),$$

where $scale_t$ is an exogenous shock facility and

$$SMM_t^D = Seasoning^D(t) * Seasonal(t) * \mathcal{F}^D(Savings_t(\cdot))$$

$$SMM_t^R = \mathcal{F}^B\left(\frac{Survival_t}{Survival_t^\mu}\right) * Seasoning^R(t) * \{\mathcal{F}^S(C_t - I_t) * \mathcal{F}_\downarrow^R(\cdot) + \mathcal{F}_\uparrow^R(\cdot)\}$$

A separate seasoning component has been added to the refinancing term as a mechanism for potential decay in refinancing behavior. In practice, the burnout component is viewed as much less important for short ARM products than fixed rate products, and the refinancing seasoning component may be used in lieu of the complex burnout function.²⁰

9 Hybrid Mortgage Prepayments

Hybrid ARMs offer the challenge of "merging" fixed rate and adjustable rate models in a consistent manner. Consider the common spectrum of hybrids: 3/1, 5/1, 7/1, and 10/1 loans. It is probably safe to assume that the 3/1 borrower is an ARM borrower, having similar sensitivity to the level and slope of the mortgage curve and a relatively short expected tenure in the home.²¹ On the other end of the spectrum, the 10/1 borrower probably differs little from a 30 year fixed rate borrower. Between the 3/1 and 10/1 behaviors are likely those borrowers who might have selected balloons in years past, but now find the 5/1 and 7/1 products more attractive. These borrowers are likely (on average) to have short initial expected tenures, but may extend their evaluation horizon²² as the reset draws nearer.

To model 5/1, 7/1, and 10/1 hybrid prepayments, the ARM specification from the previous section is (hopefully) too complex. Instead, return to the fixed rate model with some notable extensions. Define

$$\begin{aligned} C_t &\equiv \text{current hybrid reference interest rate, e.g. 5YCMS.} \\ S_t &\equiv \text{current hybrid reference spread.} \\ E_t &\equiv \text{Borrower evaluation horizon at time } t. \\ \bar{G}_t(t, t + E_t) &\equiv \text{average contractual loan rate over } E_t. \end{aligned}$$

A hybrid borrower is assumed to focus on a refinancing alternative which is on the short end of the curve: after all, this is where the borrower has positioned the existing loan. This alternative instrument has the rate $C_t + S_t$ as described for the FRM model, but the reference point on the curve as well as the spread would reflect a hybrid or balloon loan instead of an FRM.²³

²⁰In other words, assume a constant $\mathcal{F}^B(\cdot) = 1.0$.

²¹For now, 3/1 ARM's are addressed using the ARM specification from the previous section.

²²This is the time period over which the economic (dis)incentive for refinancing is considered. For these borrowers, reset risk must be incorporated in some fashion. See the appendix for additional discussion.

²³In other words, the reference rate for a 7/1 is not the 10YCMS rate but is a shorter tenor such as 5YCMS.

As with ARM's, the rate to be compared with the market alternative is not a static quantity: some assumption must be made regarding the borrower's expectations regarding loan rate reset(s). For simplicity, it is assumed that the borrower averages the monthly rate paid over the evaluation horizon E_t and uses this composite rate $\bar{G}_t(t, t + E_t)$ to calculate savings from refinancing. The loan rate resets are calculated, utilizing all relevant cap, floor and margin features, assuming the underlying rate index remains at its current spot level over the entire horizon.

10 Balloon Mortgage Prepayments

Balloon products are addressed in the same fashion as hybrids, for the balloon borrower effectively faces a "market reset" at the balloon date. Most balloon loans include an option to extend the term of the loan at a new, slightly above-market rate, and borrowers are certainly able to seek a new loan from another lender.

Perhaps the primary economic difference between the balloon and hybrid loan at the reset is the uncapped nature of the balloon relative to the hybrid. This may cause a "panic effect" as the balloon borrower approaches the balloon date, where the borrower rushes to lock in a new rate rather than bear the reset risk.

Specifically, balloons are modeled as hybrids with a notable exception: $\bar{G}_t(t, t + E_t)$ is calculated assuming a reset into a new fixed rate balloon loan at the balloon date.

11 Handling Prepayment Penalties

Prepayment penalty features can be added to any of the aforementioned products via the use of two common model components: the demographic/turnover seasoning function and the refinancing hurdle.

In particular, one would expect a selection bias in a pool of borrowers that accepted three year prepayment penalties: the expected tenure in the home is greater than three years. Thus, the usual PSA-like seasoning ramp discussed earlier would need to be replaced with a multi-tiered function which allows for a much lower turnover rate during the penalty window, but shifts to a "normal" turnover rate thereafter.

Should rates fall significantly, however, the potential savings from refinancing could exceed the amount of the penalty. To allow for rational borrower behavior in this instance, the refinancing cost hurdle is modified to include the amount of the penalty during the penalty window. As the end of the penalty window draws near, it is possible that a dampening in such refinancings would occur in order to avoid the penalty, with pent-up demand being released as a spike at the end of the penalty period.

12 Idiosyncratic Spikes

In a number of contexts, prepayment spikes are observed that may not fit cleanly into this (or any) prepayment specification. Many ARM pools show spikes immediately following resets: the unsustained nature of the surges makes them difficult to explain economically. Is the behavior a function of teaser/equity takeout plays, payment shock, or short tenure selection bias? Since ARM products explicitly include a reset to market, excluding teaser and cap effects, it seems counterintuitive that such spikes

could be pervasive.

This spike effect, as well as the "balloon panic" and "post-penalty surges" of the previous sections, can be addressed using the demographic and refinancing seasoning components of the models. Prior to incorporating such behaviors, however, some additional empirical work should be done to attempt an explanation of the magnitude and persistence of these effects.

13 Calibration and Parameter Estimation

13.1 Implied Estimates

In the absence of historical data, QF has calibrated models to provide reasonable and theoretically appealing projections that fit market observations of prices or speeds wherever possible. Results from such calibrations have been subjected to a battery of tests which stress the model performance under normal and extreme scenarios. Model performance is also examined with the goal of reconciling behavioral differences among products from different markets.

Implied calibration of the model can be performed in a number of ways.

13.1.1 Market Prices

QF's first pass at implied calibration used market prices for TBA's, IO's and PO's for a particular mortgage type (e.g. FNCL) and attempted to establish parameters that performed well across interest rate scenarios and coupons. This technique is expensive computationally, and QF abandoned the approach a number of years ago after observing that similar answers were produced from the less costly approaches that follow.

13.1.2 Third Party Models

Examination of models from certain third party vendors from which we obtained licenses or may observe via the Bloomberg, notably Andrew Davidson and Espiel, provides a range of opinions upon which to benchmark QF results. As QF has some experience with these vendors, use of the projections is not without caution, for such models have limitations and biases that QF seeks to avoid.

13.1.3 Bloomberg Dealer Projections

For generic TBA's, Bloomberg makes available PSA projections from many dealers for fixed rate mortgages of across coupons and vintages. Using such data enables QF to produce a matrix of target speeds by product type for specific coupons and interest rate scenarios. Since the mortgage issues are not completely generic in terms of WAC and WAM and the reporting date varies, some reconciliation of differences between dealers is required.

QF examines the performance of internal models relative to the dealer projections on a monthly basis. Re-calibration occurs when the QF models begin to appear extreme relative to the universe of dealer opinions: QF typically positions its model relative to other models such that MSR or IO is conservatively valued.

13.1.4 Prepay Tuner

A recent addition to the QF toolbox, the *Prepay Tuner* is a program that will measure the performance of any QF model versus actual prepay history for a specific pool. Developed to address backtesting and statistical fit issues, the output of this program will track actual balance and SMMs against model forecasted balance and SMMs on a period by period basis. In addition, it is possible to output at each period the full state of the prepay model (i.e. the value of all variables which contribute to the final forecasted SMM). Results hinge on the amount of actual data for each pool, specifically the number of loans per pool, and the number of actual vintages (coupon and vintage combinations) for a particular model type. This output gives the ability to examine exactly how each individual piece of the model performs across vintages and coupons, yielding measures of fit and potential calibration aids.

13.2 Statistical Estimation of Fixed Rate Mortgage Prepayment Models

The following is an introduction to the current Quantitative Finance statistical modeling efforts. Currently the results of statistical work are used to provide insights leading to adjustments to the implied model. Over time, it is expected that this work will form a significant portion of the modelling effort.

13.2.1 Prepayment Model Structure

$$CPR^D(t) = \text{Baseline} * \text{Lockin}(\text{Savings}) * \text{seasoning}(t) * \text{seasonality}(\text{month})$$

$$CPR^R(t) = \text{RefiIncentive}(\text{Savings}) * \text{Burnout}(t)$$

$$CPR(t) = CPR^D(t) + CPR^R(t)$$

13.2.2 Overview

In general QF assumes two main reasons for prepayments on mortgages, Housing Turnover ($CPR^D(t)$) and Refinancings ($CPR^R(t)$). Each of these sub-models is made up of interactions between various sub-components, where each sub-component is a nonlinear function of its underlying variable.

The turnover related sub-model ($CPR^D(t)$) is produced out of the interaction between four separate nonlinear sub-components: Baseline (base turnover related prepayment speed for seasoned pools), seasoning (slower prepayment speeds early in the life of the pool), Seasonality (faster prepayment speeds in warmer months), and Lockin (slower prepayment speeds in rising rate environments).

The Refinancing sub-model ($CPR^R(t)$) is produced out of the interaction between two sub-components: RefiIncentive (faster prepayments in falling rate environments), and Burnout (slower prepayment speeds during periods of refi-incentive due to changes in pool composition).

The model described here differs a good bit from typical statistical models in its functional form and in the way it is estimated; it also has a fair number of "parts." As a result, implementing a model of this type requires solutions to numerous small problems, many of which are not standard in statistical literature. Because of these restrictions, the estimation approach we use is non-parametric in nature. This approach provides the flexibility to fit the nonlinear relationships seen in the model and allows modelling of the complex interactions seen between economic refinancing and turnover-related refinancing.

13.2.3 The Essence of the Modeling Approach

As can be seen from the model, many interactions exist between variables that must be estimated in a consistent manner, which both isolates the effect of these interactions and models them correctly. For example the seasoning effect takes the form of faster prepaids in warmer months but slower prepaids for younger pools. If these two effects were modeled without careful consideration to this interaction, a blending of these two effects would occur in our estimations and biases would undoubtedly show up in the model. Because of the complicated nature of the model these types of interactions are seen often and form the most difficult part of the modeling process. To make things even more difficult these interactions are non-linear in nature, which makes them difficult to model using standard linear regression.

To address these difficulties QF has developed a modelling approach that addresses these interactions in a consistent and accurate manner. The modeling approach consists of three steps that applies to each sub-component of each sub-model. These three steps are as follows:

1. Isolate a subset of the data on which to fit each sub-component of each sub-model. Subsets should be chosen to isolate the effect of that one predictor.
2. Use nonparametric regression to model this relation. We initially use loess to do this for its flexibility and robustness with respect to outliers
3. Subtract out (or divide out) the predicted (CPR) effects of this predictor from the rest of the data.

To illustrate this process, consider again the problem of estimating the seasoning process. If one were to assume the seasoning process consisted of only the interaction between the seasoning ramp (slower prepaids for young pools) and seasonality (faster prepaids in warmer months), one would first isolate the seasonality effect by estimating it on the subset of data where the seasoning ramp has no effect, i.e. fully seasoned pools. One could then remove the seasonality effect from the entire data set by dividing out its predicted CPR. With the seasonality effect removed it is possible to estimate the entire seasoning ramp without concern for the possible confounding effects of the seasonality variable.

As can be seen from the example above, the order in which sub-components are estimated is important. For instance, a decision to estimate the seasoning ramp before the seasonality multiplier would require finding a portion of the data in which the seasonality multiplier was constant. This is obviously much more difficult to do (if not impossible) than estimating in the reverse direction.

13.2.4 Modeling Specifics

At the coarsest level, observations are divided into two types: demographic (data without refi incentives) and refi (data with refi incentives). This division is based on the values of the savings variable found in the model. This variable is a function of rate differential, loan size, borrower evaluation horizon and assumed transaction costs, and is used throughout the model as a measure of economic incentive or disincentive to prepay.

The demographic sub-model is fitted first, using only demographic observations, for it is possible to find data in which to isolate demographic prepayments, i.e., loans which prepay when they have no rate incentive to do so. It would not be possible to isolate groups of loans known to be solely refi-prepayments. A predicted demographic baseline prepayment rate is then subtracted out from the refi

observations, so what remains in the refi observations is just prepayments due to refi. The refi sub-model is then estimated.

Within the demographic sub-model, a baseline component is fitted first, the seasonality sub-component second, then the rate-sensitive component (Lockin), and finally the seasoning component:

Seasonality is estimated in one of two ways: (1) pre-specified by National Association of Realtors ratios; (2) estimated from the data. In the event it is estimated from the data, it should be estimated ONLY on the subset of the demographic data that is fully seasoned (and those which are not too far out of the money), to avoid the confounding effects of the other predictors. In other words, for loans that are fully seasoned, CPR does not vary much with Seasoning, and for loans that are demographic data but not deep out of the money, CPR does not vary much with Savings, so QF has isolated a subset of data for which Seasonality is the driving force behind CPR for the model specification. Call this subset "Seasonality data."

Once CPR as a function of Seasonality is fitted, its effects are removed from both the demographic data set and the Seasonality data set, so remaining variability is due to other predictors. CPR as a function of Savings (Lockin) is then estimated on the residuals of the demographic data, and its effects are removed, so remaining variability in CPR in the demographic data set is due to Seasoning. Finally, CPR as a function of age(Seasoning) is estimated on the residuals of the demographic data set from the previous step to complete the demographic sub-model estimation.

Next, the predicted CPR is determined on the Refi data set using the demographic sub-model. This Demographic-predicted CPR is subtracted from the actual CPR; use the name CPRrefi to describe the residual CPR that remains. Because the effects of the demographic sub-model have been subtracted out, that remaining CPR is assumed due to economic refinancing.

Within the Refinancing sub-model, effects due to Savings are estimated first, and then the effects due to Burnout. This process begins with isolation of a subset of the data which has seen only moderate amounts of economic related refinancing opportunities. Call this the refi-savings data set or the unburned data set. CPR is estimated as a function of savings (refi) on this data set. Predicted CPR is then estimated as a function of savings on the entire Refi data set, and its effects are removed.

The Burnout multiplier is then estimated on residuals attained from the previous step on the entire Refi data set to complete the estimation process.

Results of all nonparametric regression fits are converted to a lookup table between dependent and independent variables. A finely spaced grid of values for each independent variable is used to create the lookup table, so intermediate values can be determined via linear interpolation.

13.2.5 Details on Burnout

Burnout is defined as the decrease in interest rate sensitivity of a pool as more and more refinancing opportunities are experienced. QF measures the burnout of a pool by comparing the predicted prepayments due to refinancings with the predicted prepayments due to turnover. This measure is internally consistent and its range is always contained in [0,1]. The measure takes the form:

$$Burnout = \frac{survival_t^{D+R}}{survival_t^D}$$

Where

$$\begin{aligned} survival^{D+R} &= \prod_t [1 - (SMM_t^D + SMM_t^R)] \\ survival^D &= \prod_t [1 - SMM_t^D] \end{aligned}$$

This measure of Burnout is an improvement over the often-used pool factor measure. The reason for this is within the pool factor framework it is difficult to designate between burnout measures of older low-coupon pools that have never seen refinancing opportunities but have a small pool factor due to turnover related prepayments alone, and younger pools that have experienced substantial refinancing opportunities. This new Burnout measure eliminates the possible confusion.

13.2.6 A Brief Primer on Local Regression (loess)

Regular regression, also known as ordinary least squares (OLS) regression, determines the best-fitting straight line between a dependent variable y and an independent or predictor variable x by minimizing the sum of squared errors between the predicted values of y and the actual values of y . This results in the following estimates:

$$(y - \bar{y}) = \beta(x - \bar{x}) + \varepsilon$$

where β is the estimated regression coefficient. Another fundamental measure of the relationship between two variables x and y is the coefficient of correlation r . This measure takes the form:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(ns_x s_y)}$$

where s_x and s_y are the sample standard deviations of x and y .

Given mean-centered x and y , the connection between correlation and regression is

$$\beta = \left(\frac{s_y}{s_x} \right) r \tag{1}$$

So one can think of the correlation as a standardized slope (regression beta), or one can think of the beta as a scaled correlation.

Changing from sample statistics to population parameters, the previous relation can be expressed as

$$\rho = \frac{\beta \sigma_x}{\sigma_y}$$

or

$$\rho^2 = \frac{\beta^2 \sigma_x^2}{\sigma_y^2}.$$

And because

$$y = \beta x + \varepsilon,$$

it follows that

$$\rho^2 = \frac{\beta^2 \sigma_x^2}{\beta^2 \sigma_x^2 + \sigma_\varepsilon^2}$$

Local correlation is simply a generalization of this relation in that betas and correlations are allowed to vary as a function of independent variable x , instead of being constant for all x (note that one can also generalize the variance terms to be non-constant as well):

$$\rho^2 = \frac{\beta^2(x) \sigma_x^2}{\beta^2(x) \sigma_x^2 + \sigma_\varepsilon^2}$$

For an intuitive overview of this topic, see Blythe (1994; more technical references are also given there).

The name Loess stands for Local regression. Essentially, local regression or loess models are used when the slope or correlation between two variables is not constant across the values of the predictor (or their variances are not constant). One can think of loess as many different regressions over different locally weighted "windows" in the data set.

Cleveland (Visualizing Data, pp.91-101, 110-119, 122-127) provides the technical details of loess, including:

1. How loess is made to be robust with respect to outliers,
2. Quadratic versions of loess (i.e., a local version of quadratic regression) and
3. Choice of smoothing parameters in loess.

A Appendix: Economic Value of Refinancing Fixed Rate Mortgages

A.1 Equations

Define

- $G \equiv$ contractual mortgage coupon rate.
- $M \equiv$ market mortgage coupon rate.
- $T \equiv$ maturity of existing and prospective loans in months.
- $E \equiv$ borrowers evaluation horizon in months.
- $U \equiv$ current loan balance.

The standard fixed rate payment calculation is

$$P(U, M, T) = U * \left(\frac{\frac{M}{12}}{1 - \frac{1}{(1 + \frac{M}{12})^T}} \right).$$

Define

$$\alpha(M, T) = \frac{1}{1 - \frac{1}{(1 + \frac{M}{12})^T}}.$$

Thus,

$$P(U, M, T) = U * \left(\frac{M}{12} \right) * \alpha(M, T).$$

Assume a borrower evaluates the decision to refinance over some horizon $E < T$. Thus, prior to considering costs,

$$\text{Savings} \equiv U * \sum_{i=1}^E \left[\frac{(\alpha(G, T) \frac{G}{12} - \alpha(M, T) \frac{M}{12})}{(1 + \frac{M}{12})^i} \right] \quad (2)$$

$$= \frac{U}{12} * [\alpha(G, T)G - \alpha(M, T)M] * \sum_{i=1}^E \left(\frac{1}{(1 + \frac{M}{12})^i} \right) \quad (3)$$

$$= \frac{U}{12} * [\alpha(G, T)G - \alpha(M, T)M] * \left(\frac{\alpha(M, E)^{-1}}{\frac{M}{12}} \right) \quad (4)$$

$$= U * \left[\frac{\alpha(G, T) G}{\alpha(M, E) M} - \frac{\alpha(M, T)}{\alpha(M, E)} \right]. \quad (5)$$

A.2 Coupon Ratio versus Differential

Early prepayment models used a coupon differential ($G - M$) as the variable which explains refinancing. An examination of equation (4) pointed subsequent modelers toward a coupon ratio $\frac{G}{M}$ instead, as level

effects are clearly present.²⁴ From a purely mathematical perspective, neither proxy is perfect: for a given coupon differential ($G - M$), the savings from refinancing given in (4) changes smoothly as the level of rates (M) approaches zero. The ($G - M$) differential is static, whereas the ratio $\frac{G}{M}$ explodes as rates get extremely low.²⁵ Furthermore, the $\frac{G}{M}$ approximation worsens as maturity T gets smaller or as the horizon E is assumed to be significantly shorter than T .

These conclusions imply that the choice of ratio versus differential might vary by product type, borrower demographics, and seasoning. Such complexity suggests that the use of (4) directly, with its natural inclusion of loan size and level effects, makes more sense than either proxy.

A.3 Modeling the Borrower Evaluation Horizon

It seems unlikely that the majority of borrowers truly intend to remain in a home for the duration of the associated loans. Furthermore, for the products favored by "fast movers," a pool of borrowers will likely have an average evaluation horizon which changes over time. In order to incorporate these effects, consider the following framework:

- $L \equiv$ borrowers anticipated dwelling time at origination in months.
- $E_t \equiv$ borrowers evaluation horizon in months, $E_t \leq L$.
- $MaxE \equiv$ maximum evaluation horizon in months.
- $MinE \equiv$ minimum evaluation horizon in months.
- $MaxAt \equiv$ time where $MaxE$ is reached in months.

Consider a 7-1 borrower: assume the initial anticipated dwelling time $L = 60$ months. As the borrower may see some uncertainty in the forecast of L , set $E_0 = 48$ months. Assume $MinE = 12$ months, $MaxE = 60$ months, and $MaxAt = 84$ months, the time of the first coupon reset. Now let E_t evolve as follows:

$$E_t = E_0 \quad \forall t \in [0, L - E_0] \quad (6)$$

$$E_t = L - t \quad \forall t \in [L - E_0, L - MinE] \quad (7)$$

$$E_t = MinE + \frac{MaxE - MinE}{MaxAt + MinE - L} * (t + MinE - L) \quad \forall t \in [L - MinE, MaxAt] \quad (8)$$

$$E_t = MaxE \quad \forall t \geq MaxAt \quad (9)$$

In the case of the 7-1 borrower, what do these equations imply? Initially, the borrower expects to live in the home for 5 years (L), but will consider refinancing if the costs are exceeded in 4 years (E_0). For the first year of the loan ($L - E_0 = 12$), the borrower continues to use a four year evaluation horizon. After this first year, however, the evaluation horizon begins to fall, for the borrower begins to approach the original dwelling forecast L ; the expected dwelling time remaining in the home is falling, so the time period for refinancing cost recovery falls also. Equation 6 assumes this reduction is linear until the minimum evaluation horizon $MinE$ is reached. At this point, the evaluation horizon begins increasing again, reflecting the possibility that the original L was underestimated, and the borrower may remain in

²⁴Note that for a given coupon reduction (say 1 percent), the borrower sees greater absolute payment savings (undiscounted) at higher rate levels. Equation (4) suggests the opposite effect, that savings will be higher at lower rate levels due to the reduction in discounting. However, this crossover only occurs if the borrowers' horizon E is long enough.

²⁵Stochastic interest rate models inevitably produce such scenarios.

the home beyond five years. The horizon continues to expand until the initial reset date, where it reaches a maximum: the borrower is now facing a market reset originally unanticipated, and the model assumes that the borrower will seek a fixed rate alternative using an evaluation horizon of material length.

The framework is fairly complex when used to address hybrid ARM's and balloons, where there are elements of fixed rate behaviors which transition to ARM behaviors. For pure fixed rate and ARM products, the implementation is drastically simplified such that $E_t = \bar{E}$. Note that in the context of a pool of mortgages, some dispersion in the horizon across borrowers will be naturally incorporated in the burnout component.

A.4 Points Effects: Above Market Originations

Consider the loan rate at origination, G_0 . Using lag structure built into the model already, observe the market reference rate, M_0 .

$$X = G_0 - M_0, \quad (10)$$

the amount by which the loan is "above market."

To address this: Record X . Create

$$X' = \min(X, \text{Upper bound}), \text{ if } X \geq \text{Lower bound} \quad (11)$$

$$X' = 0 \text{ if } X < \text{Lower bound} \quad (12)$$

X' is therefore constrained to a range controlled by the modeler. Assume, for example,

$$\text{Lower bound} = 25bp$$

$$\text{Upper bound} = 100bp$$

Normally, refinancing is driven as a function of G_0 vs. M_t . Now adjust this:

$$M'_t = M_t + X' \cdot \alpha(t), \quad (13)$$

where $\alpha(t)$ might look like the following:

$$\alpha(0) = 1$$

$$\delta_1 = 0.02$$

$$\delta_2 = 0.10$$

$$\alpha(t) = \alpha(t-1) \cdot (1 - \delta_1) \quad \forall t \leq 12$$

$$\alpha(t) = \alpha(t-1) \cdot (1 - \delta_2) \quad \forall t > 12$$

$$\alpha(t) = 0 \quad \forall t > 60$$

Thus the above market effect would diminish slightly over year 1, rapidly over years 2-3, and is virtually 0 thereafter.

A.5 Implementation Issues

Included below are details on the current production implementation of the model.

- **Refinancing Vehicle Assumptions:** The assumption that a borrower always refinances into a mortgage of matching maturity simplifies the problem considerably, as significant mismatches in current maturity and new loan maturity would introduce material differences in principal amortization. Note that under the simplifying assumption, seasoned pools would need a reference refinancing vehicle which is shorter than the original WAM of the pool.
- **Lags:** For historical reference rates (i.e. dates prior to the pricing date), the model looks back a period of d days to observe a rate. That rate is the average reference rate over a period of m previous months. Both d and m are specified with the prepay model parameters and have values of 15 and 1, respectively.
- **Spreads:** The use of spreads is aimed at creating mortgage rates from swap rates within the Monte Carlo model. Where applicable the model compares the product coupon to a reference rate (such as 10Y CMS) and computes a spread between the two. The mean of that spread is calculated over a specific number of months, typically 24. The spread at pricing date is also computed: currently a 5 day average is used to reduce noise. The speed of reversion γ for the spread mean is 0.33333.

In the case of jumbo products or others where the relevant reference rate might be the rate on a similar conforming product plus a spread, that additional spread is included with the prepay model parameters.

- **Horizons:** As noted above, savings for the demographic and refi incentive components are calculated based on horizons specified with the model parameters. The current implementation of the model does not make full use of the separate horizon features: horizons are set at 60 months.
- **Costs:** The model incorporates both fixed and variable cost hurdles of \$1,000 and 25 basis points of loan balance, respectively.
- **Loan Size:** Beyond the natural impact of loan size on Savings calculations, there is no adjustment of the model for loan size except in the reference rate, e.g., jumbos may get a positive spread to conforming production.
- **Points Effects:** Currently all parameters are set to zero.
- **Prepay Penalties:** Prepayment penalties are entered as a percent of the loan balance. The value and the number of months from origination over which the penalty is applicable are entered both in the prepay model parameters heading section and in the mortgage security file. Such mortgages are modeled as a special security type.

In order to fit the model behavior, most demographic seasoning ramps are adjusted to season fully after the prepayment penalty period terminates. Prior to that, they might typically season up to some lower level (e.g. 50%) by the end of what would be a normal seasoning period for a similar mortgage without penalties.

- Prepay Model State Information: Options are available through two pieces of code, **ppm_tuner** and **ppm0_state**, which $\forall t \in [origination, maturity]$ will output the values of each contributing variable in the prepay model. **ppm_tuner** works with historical data, and **ppm0_state** works prospectively.