# A Non-Parametric Prepayment Model and Valuation of Mortgage-Backed Securities

NARASIMHAN JEGADEESH AND XIONGWEI JU



**NARASIMHAN JEGADEESH** is the Brandt distinguished professor of finance at the University of Illinois at Urbana-Champaign.

**XIONGWEI JU** is a quantitative strategist in the division of quantitative trading at Knight Financial Products, LLC, in White Plains, New York.

Residential mortgages represent the single largest category of private debt in the U.S. market. Financial institutions that lend against home mortgages pool a large part of these loans and sell them in the capital markets as mortgage-backed securities (MBS). The MBS market has grown rapidly since the early 1980s; the current value of MBS outstanding is over $1.8 trillion. Several MBS derivative instruments, such as collateralized mortgage obligations and principal-only and interest-only securities, are also actively traded in this market. These derivatives essentially divide the cash flows from mortgagors into different slices to cater to various investor clienteles. Mortgage-backed securities and their derivatives offer investors a variety of investment opportunities and serve as an efficient channel for funds from the capital markets to homeowners.

These are complex instruments that are difficult to price. Much of the difficulty arises because residential mortgagors are typically allowed to prepay any portion of the mortgage principal outstanding at any time during the life of the mortgage. This prepayment option is the same as a call option, and hence MBS are callable bonds. The theory of pricing callable bonds cannot be directly applied to price MBS, however, because individual mortgagors do not always make the "optimal" call decisions predicted by purely financial considerations. Many mortgagors, for example, do not prepay when the value of the mortgage exceeds the value of the principal, and many other mortgagors prepay when the value of the mortgage is below the value of the principal. We have to be able to characterize mortgagors' prepayment behavior in order to determine the value of MBS and their derivatives.[1]

This article introduces a non-parametric method for estimating prepayment models. Non-parametric methods are particularly suitable in this context, since researchers have strong priors about the underlying factors that drive prepayment decisions, but parametric forms of the functional relation between these factors and prepayment probabilities are unknown. For instance, it is well known that the probability of mortgage prepayment is related to the age of the mortgage, but there is no underlying theory that guides researchers to specify this relationship parametrically. Schwartz and Torous [1989] (S&T) and Stanton [1995] impose parametric restrictions on this relation for tractability, but the log-logistic functional forms that S&T specify or Stanton's time-independence assumption are not consistent with the data. Our approach lets the functional relation be determined empirically from the data.

We adapt the generalized additive model (GAM) technique to estimate a mortgage prepayment model. This technique, first developed in a different setting by Hastie and Tibshirani [1986], is a non-parametric varia-

tion of generalized linear models. It replaces the linear relations in generalized linear models with unspecified smooth functions. The GAM consists of a sum of such functions. We adapt the GAM for the prepayment setting and implement it using a local scoring method combining transformation, reweighting, and backfitting techniques.

Our non-parametric estimation technique generalizes the multivariate density estimation (MDE) techniques used by Ait-Sahalia [1996], Boudoukh et al. [1997], and others. The MDE method is highly data-intensive, and it is virtually impossible to estimate models with more than two explanatory variables using this technique.[2]

Boudoukh et al., for example, estimate MBS prices as functions of only the level of interest rates and slope of the structure, and not of any of the prepayment characteristics of various pools because of methodological constraints. These constraints limit the applicability of the models. In the case of Boudoukh et al., the model may be applied for pricing new MBS issues prior to any prepayments, but it cannot be used either to price seasoned MBS and their derivatives or to hedge their interest rate risks.

The GAM technique allows for the covariates to be considered either jointly or individually. When the effects of the covariates are specified as separate univariate or bivariate non-linear functions, the dimensionalities of these functions do not increase exponentially as in the case of the MDE technique, and hence the data requirement does not increase exponentially with the number of factors. Additionally, the GAM technique allows for mixed parametric and non-parametric specification of the relations between the dependent variable and independent variables.

An important innovation in our work is the reformulation of the likelihood function that allows us to estimate the model using non-parametric methods. In the full likelihood function specification in earlier work (such as Green and Shoven [1986] and S&T), the contribution of each observation to the total likelihood depends not only on the covariates at the time of prepayment but also on the history of all covariates from the time of mortgage origination to the time of prepayment. This form of the likelihood function is cumbersome to estimate even with parametric assumptions, so Green and Shoven and Schwartz and Torous impose a simplifying assumption that past covariates equal the covariates at the time of prepayment. Such assumptions lead to estimation inefficiencies and yield biased parameter estimates. We reformulate the likelihood function and get it in a form which we show is equivalent to estimating a binomial generalized additive model; this equivalence enables us to apply non-parametric estimation methods.

Our approach also allows us to estimate a path-dependent prepayment model. Allowing for path-dependence is particularly important in the context of prepayments. Intuitively, the probability that a mortgagor in a pool will prepay in the next instant depends not only on whether the interest rate today is low relative to the mortgage coupon but also on whether the mortgage pool had previously experienced low interest rate environments since its origination.

We use two variables that measure the expected and unexpected levels of burnouts, given the past realizations of the covariates such as interest rates to capture such path-dependence. The expected level of burnout captures a pool's cumulative exposure to prepayment risks in the past, and the unexpected level of burnout captures the heterogeneity in prepayment speeds across pools.

Our model is estimated using a sample of prepayment data for GNMA pools of thirty-year mortgages on single-family homes over the 1971-1998 sample period. We find the relationship between prepayment rates and the variables in our model — age of the mortgage; ratio of the mortgage coupon rate and prevailing interest rates; and expected and unexpected burnouts — to be highly non-linear, so it would be difficult to capture these relations with parametric functions. Our model fits the prepayment data significantly better than the parametric models.

We also illustrate the relationships between the factors that affect prepayments and the prices of mortgage-backed securities. We find that the estimated shape of the baseline hazard function fundamentally affects the theoretical MBS prices. It is important to decompose pool burnout into expected and unexpected components, as we propose. When we consider two pools with the same level of total burnouts but with different proportions of expected and unexpected burnouts, we find that their prices in general are different.

## I. PREPAYMENT MODEL

The first step in designing a prepayment model is to identify the factors that affect prepayments. We describe the factors that affect prepayment in our model, and explain the relationship that we expect between these factors and prepayment rates.

### Measure of Prepayment

Prepayments are typically measured using single-month mortality (SMM) or conditional prepayment rates (CPR). SMM is the proportion of the outstanding value of a mortgage pool that is prepaid in a given month. CPR is the annualized prepayment rate defined as:

$$CPR = 1 - (1 - SMM)^{12}$$

In our model, the time to prepayment for each mortgagor is denoted by the random variable T, and f(t) represents its probability density function. The survival function S(t) is defined as the probability that a particular mortgage survives at least until time t, or formally $S(t) = Pr(T \geq t)$. The hazard function is defined as

$$\pi(t) = \frac{f(t)}{S(t)}$$

Intuitively, $\pi(t)dt$ is the probability that a given mortgage will be prepaid during the next instant of time dt, conditional on it being still alive at t. It is essentially a continuous-time version of CPR and SMM, and it measures instantaneous prepayment risk.

### Age of Mortgage — Baseline Hazard Function

The residential mortgages that underlie MBS typically are not assumable on home sale. Therefore, prepayment occurs whenever the mortgagor sells the home. Home sales occur for a variety of reasons such as job changes, marriages, and divorces. Many of the reasons for home sales are unrelated to interest rates, so standard interest rate-contingent claims modeling approaches cannot account for their effects on the value of the MBS.

The function that describes the relation between the age of a mortgage and the prepayment rates keeping other factors constant is called the *baseline hazard function*. A common benchmark for this function is the Public Securities Association (PSA) model. The PSA model specifies that CPR increases by 0.2% every month for the first thirty months and remains at the 6% level thereafter. This model loosely describes the empirical relation between prepayment rates and age according to historical experience, and it captures the fact that newly issued mortgages are less likely to be prepaid than seasoned mortgages.

Since we have no theory to guide our choice of functional form for the baseline hazard function, we simply specify that:

$$\pi_i(t) \propto f_a[age_i(t)]$$

where $\pi_i$ is the hazard rate for the mortgages in the i-th pool, and $f_a$ is a continuous function.

### Current Interest Rate versus Mortgage Rate — Option Effect

Mortgagors are more likely to prepay when refinancing rates are below the coupon rates on their mortgages. The fact that the mortgagors tend to exercise the prepayment option when it is in the money is by itself not surprising. What is intriguing, however, is the fact that many mortgagors do not prepay even when the refinancing rates are substantially below the rates they pay on their existing mortgages. For instance, the refinancing rates were well below 7% in 1998, and yet several pools backed by mortgages originated in the early 1980s at rates well over 10% were not yet fully prepaid.

Timmis [1985] and others propose that mortgagors may delay prepayments because of transaction costs. The extent of the financial benefits to potentially prepaying the high-coupon mortgages now in the market suggests that such transaction costs have to be fairly high, and these costs cannot be fully captured by the direct costs of processing documents for refinancing. A potential source of indirect costs that adds to the transaction costs of refinancing is the cost of acquiring information about current mortgage rates. Also, the credit quality of some of the mortgagors may have deteriorated over time, and it may be virtually impossible for these homeowners to refinance at economically attractive rates. In some cases, house prices may have fallen, which would impair the ability to refinance the existing mortgage balance (see Archer et al. [1996]).

Regardless of the sources of these indirect costs, mortgagors will prepay when the financial benefit from prepayment exceeds their private transaction costs (both direct and indirect). Given any distribution of transaction costs across homeowners, it is reasonable to expect the hazard rate to be a function of benefits from prepayment. We use the ratio c/L to capture the financial benefit to mortgage prepayment, where c is the mortgage coupon rate and L is the current long-term interest rate. We use the ratio c/L as the argument in this function rather than the difference c – L because, as Hayre and Rajan [1995]

show, the ratio is closer to being proportional to the value of the refinancing benefit.

Since there is a time lag between when a mortgagor decides to prepay and when he or she is able to close the new mortgage, we specify the hazard rate as a function of the long rate lagged three months:

$$\pi_i(t) \propto f_r[c/L(t-0.25)]$$

The functional form of $f_r$ will depend on the distribution of transaction costs across investors, and there is no good basis for parameterizing this function. Archer and Ling [1993] and Stanton [1995] assume particular distributions of transaction costs across mortgagors, and derive structural relations between hazard rates and interest rates. Because of the flexibility of our approach, we are in a position to let the data tell us the relation between the hazard rates and the refinancing benefits without imposing additional assumptions.

**Burnout Effect**

With the passage of time, a certain proportion of mortgage pools prepay the outstanding principal and leave the pool. The conventional wisdom is that the mortgagors who leave the pool early are those who are more sensitive to changes in interest rates or who have lower refinancing costs, and that homeowners remaining in the pool are typically less interest rate-sensitive or have higher refinancing costs than the ones who prepaid.

The literature therefore specifies the hazard rate as a function of pool burnout BO(t), defined as:

$$BO(t) = 1 - \frac{AO(t)}{AO_0(t)}$$

where AO(t) is the total principal outstanding at time t, and $AO_0(t)$ is the principal that would have been outstanding in absence of prepayment.[3]

The arguments suggest a "burnout hypothesis" that the hazard rate will be a decreasing function of burnout. Consider, however, a scenario where refinancing rates have been higher than the mortgage coupon rate since origination. Since it is costly to prepay mortgages after an increase in interest rates, any burnout in such a pool will be due to the departure of relatively less interest rate-sensitive homeowners. Therefore, the relation between hazard rate at time t and burnout depends not only on the level of burnout but also on the path of interest rates between the date of mortgage origination and time t.

To incorporate such path-dependence in prepayments, we decompose burnout into expected and unexpected components. The expected burnout component $\overline{BO}(t)$ is defined as:

$$\overline{BO}(t) = 1 - \frac{\overline{AO(t)}}{AO_0(t)}$$

where $\overline{AO(t)}$ is the expected principal outstanding according to our model, conditional on the history of the pool. If the pool has been exposed to high interest rate environments in the past, the expected burnout will be low. Conversely, if the pool has been exposed to low interest rate environments, the expected burnout will be high. We expect a negative relation between expected burnout and hazard rates.

Therefore:

$$\pi_i(t) \propto f_{\overline{BO}}\left[\overline{BO_i}(t)\right]$$

Mortgage pools typically consist of mortgages originated within contiguous geographic areas. Since demographics vary across different geographic areas, there are likely to be some pool-specific variations in prepayment rates. For example, a pool with a higher-than-average proportion of highly mobile population, such as a Silicon Valley pool, is likely to prepay at a faster rate over its entire life than a pool in a geographic location with less mobile population and low housing turnovers. We use "unexpected burnout" to take into account such heterogeneity across pools.

Unexpected burnout $\widetilde{BO}(t)$ is defined as:

$$\widetilde{BO}(t) = BO(t) - \overline{BO}(t)$$

Pools with high unexpected burnout are fast prepayment pools, and we expect a positive relation between this variable and the hazard rate. The relation between the hazard rate and $\widetilde{BO}_i(t)$ in our model is:

$$\pi_i(t) \propto f_{\widetilde{BO}}[\widetilde{BO}_i(t)]$$

### Seasonality

Homeowner relocations tend to be more concentrated in summer months than in other months because of schooling considerations. As a result, housing turnover and mortgage prepayments also exhibit a summer seasonality. To account for this seasonality, we include a summer dummy in our model, defined as:

$$\begin{cases} D(t) = 1 & \text{if t = May to August} \\ D(t) = 0 & \text{Otherwise} \end{cases}$$

The relation between the summer dummy and the hazard rate becomes:

$$\pi_i(t) \propto e^{\beta D(t)}$$

We expect $\beta > 0$ because more mortgagors prepay in the summer.

### Combined Effect of Factors

Combining the factors, we can write the prepayment hazard function for the mortgages in the i-th pool at time t as

$$\pi_i(t) \propto f_a[age_i(t)] f_r[c_i / L(t-0.25)] \times$$

$$f_{\overline{BO}}\left[\overline{BO}_i(t)\right] f_{\widetilde{BO}}[\widetilde{BO}_i(t)] e^{\beta D(t)}$$

or

$$\ln \pi_i(t) = \pi_0 + \ln f_a[age_i(t)] + \ln f_r[c_i / L(t-0.25)] +$$

$$\ln f_{\overline{BO}}\left[\overline{BO}_i(t)\right] + \ln f_{\widetilde{BO}}[\widetilde{BO}_i(t)] + \beta D(t)$$

where $\pi_0$ is the constant of proportionality.

We can rewrite this expression as:

$$\ln \pi_i(t) = \pi_0 + f_0[age_i(t)] + f_1[c_i / L(t-0.25)] +$$

$$f_2\left[\overline{BO}_i(t)\right] + f_3[\widetilde{BO}_i(t)] + \beta D(t)$$

The covariates that determine the hazard rate are $v = (age, L(t-0.25), \overline{BO}, \widetilde{BO}, D_t)$, and the parameters and functions that need to be estimated are $\theta = (\pi_0, f_0, f_1, f_2, f_3, \beta)$. We impose no restrictions on the functions f other than continuity.

## II. ESTIMATION

### The Full Likelihood Function

Given the specification of the prepayment hazard, $\pi[t; \theta, v(t)]$ at time t, we can write the expressions for the survival function, which gives the probability that a given mortgage will be alive at time t, conditional on the history of covariates.

The survival function and the hazard function have the relationship:

$$S(t; \theta, \bar{v}) = \exp\left(-\int_0^t \pi[\tau; \theta, v(\tau)] d\tau\right) \quad (1)$$

where $\bar{v}$ is the time series of v from time 0 to t.

The probability density function of the time to prepayment is

$$f(t; \theta, \bar{v}) = S(t; \theta, \bar{v}) \pi[t; \theta, v(t)]$$

If the prepayment time for mortgage j in pool i is observed at $t = T_{ij}$, the contribution of this observation to the total likelihood function is $f(T; \theta, \bar{v}) = S[T; \theta, \bar{v}) \pi(T; \theta, v(t)]$. Some mortgage pools have left-censored data; i.e., we know the proportion of mortgages in a pool that are prepaid prior to t, but we do not know exactly when they prepaid. The contribution of left-censored observations to the likelihood function is $(1 - S[T; \theta, \bar{v})]$. If the observation is right-censored (i.e., we know only that the mortgage had not been prepaid at time T), its contribution to the likelihood function is $S(T; \theta, \bar{v})$. Finally, if the observation is both left- and right-censored (if we know only that the mortgage was prepaid between time interval $\underline{T}$ and $\overline{T}$), then its contribution is $S(\underline{T}; \theta, \bar{v}) - S(\overline{T}; \theta, \bar{v})$.

Therefore, the total likelihood function for a sample of prepayment data is:

$$L(\theta) = \prod_{mortgage_{ij} \in M_1} [1 - S(T_{ij}; \bar{v}_{ij}, \theta)] \times$$

$$\prod_{mortgage_{ij} \in M_2} [\pi(T_{ij}; v_{ij}(T), \theta) S(T_{ij}; \bar{v}_{ij}, \theta)] \times$$

$$\prod_{mortgage_{ij} \in M_3} (S(T_{ij}; \bar{v}_{ij}, \theta) \times$$

$$\prod_{mortgage_{ij} \in M_4} [S(\underline{T}_{ij}; \bar{v}_{ij}, \theta) - S(\overline{T}_{ij}; \bar{v}_{ij}, \theta)] \quad (2)$$

where $M_1$ to $M_4$ denote the sets for mortgages with left-censoring, no censoring, right-censoring, and both left- and right-censoring.[4]

Substituting Equation (1) into Equation (2), we get a likelihood function that depends solely on hazard functions:

$$
\begin{aligned}
L(\theta) = & \prod_{\text{mortgage}_{ij} \in M_1} [1 - \exp(-\int_{D_{ij}}^{T_{ij}} \pi[\tau; \theta, v(\tau)]d\tau)] \times \\
& \prod_{\text{mortgage}_{ij} \in M_2} [\pi(T_{ij}; v_{ij}(T), \theta) \exp(-\int_{D_{ij}}^{T_{ij}} \pi[\tau; \theta, v(\tau)]d\tau)] \times \\
& \prod_{\text{mortgage}_{ij} \in M_3} \exp(-\int_{D_{ij}}^{T_{ij}} \pi[\tau; \theta, v(\tau)]d\tau) \times \\
& \prod_{\text{mortgage}_{ij} \in M_4} [\exp(-\int_{D_{ij}}^{T_{ij}} \pi[\tau; \theta, v(\tau)]d\tau) - \\
& \qquad \exp(-\int_{D_{ij}}^{\overline{T}_{ij}} \pi[\tau; \theta, v(\tau)]d\tau)]
\end{aligned}
\tag{3}
$$

where $D_{ij}$ is the time of the issuance of the mortgage j in pool i. Although there are minor variations in origination dates, we will assume that all mortgages in a given pool originated at the same time because our database does not provide origination dates for individual mortgages.

The likelihood function (3) is specified in terms of integrals of historical hazard rates for each mortgage. Because all covariates change with time, computations of these integrals require that we observe the covariates at each mortgage's prepayment date and at every point during the mortgage history. This time-dependence of covariates is a common source of problems in estimating the likelihood functions in survival analysis.

It is important to understand the distinction between the terms *path-dependence* and *time-dependence*. Time-dependence is present in almost all proportional hazard prepayment models including ours. Time-dependence affects the likelihood function because the likelihood of a mortgage prepaying at time T depends on the probability that the mortgage survives until time T, which depends in turn on historical covariates. When the hazard rate at time T depends on the past path of covariates, we refer to it as path-dependence. The relation between $\overline{BO}$ and $\widetilde{BO}$ and hazard rates that we incorporate in our model gives rise to path-dependence. The extant models in the literature do not incorporate path-dependent covariates, but their likelihood functions are exposed to time-dependence.

The integrals in the likelihood function are usually cumbersome because the number of terms can explode as $T_{ij}$ becomes large. The computational complexity is on the order of the square of $T_{ij}$. Schwartz and Torous acknowledge this problem, and state:

> It is important to recognize that the assumed prepayment function involves time-varying covariates…. In general, the entire path of a time-varying covariate influences the probability of payment… [but] for empirical tractability we follow Green and Shoven (1986) and assume that a mortgagor considers only current value of the covariates, as opposed to their past or future values [1989, p. 381].

Thus S&T do not base their estimation on the full likelihood function, and they ignore the time-dependence of the covariates. In effect, they assume that the survival function at time t is given by

$$S(t) = \exp[-\exp[v(t)\theta]\int_0^t \pi_0(\tau; \gamma, p)d\tau]$$

while the correct survival function is

$$S(t) = \exp[-\int_0^t \exp[v(\tau)\theta]\pi_0(\tau; \gamma, p)d\tau]$$

Ignoring time-dependence introduces biases and diminishes efficiency. Since mortgages are more likely to be prepaid when the covariates are more conducive to prepayments, assuming that the past covariates equal the covariates at prepayment time will overestimate past hazard functions, and therefore underestimate the survival function. Also, the estimation does not take advantage of the information that a mortgage that prepaid at time T did not prepay for a particular history of covariates. Ignoring information in the data typically leads to loss of estimation efficiency, and indeed, as we show, this loss of efficiency is non-trivial.

### Reformulation of Likelihood Function

We adopt a different approach and reformulate the likelihood function so that we can take account of this time-dependence. Our reformulation of the likelihood function also allows us to include path-dependent covariates. Our data, as well as the data in most other prepayment

model research, are monthly prepayment data for pools. Therefore, for each pool i, we have a vector of sampling times $[T_{i1}, T_{i2}, T_{i3}, \ldots, T_{iM_i}]$ that are associated with the sampling period of the prepayment observations, where $M_i$ is the number of sampling points for pool i.

Suppose that at the beginning of each sampling interval $T_{ik}$, $R_{ik}$, is the number of mortgages in pool i that remain outstanding. Then $P_{ik} = R_{ik} - R_{ik+1}$ is the number of mortgages that are prepaid in the time interval $(T_{ik} - T_{ik+1}]$, and the contribution to the likelihood function from observations in pool i over this time interval is $[S(T_{ik}; \bar{v}_{ik}, \theta) - S(T_{ik+1}; \bar{v}_{ik+1}, \theta)]^{P_{ik}}$.

Notice here that we treat all the prepayments as two-side-censored. During the time interval $(T_{ik} - T_{ik+1}]$, we know there are a total of $P_{ik}$ mortgages prepaid, but we don't know the exact time of the prepayments. The mortgages that are still outstanding after the end of observation time $T_{i,M_i}$ are treated as right-censored.

Now, we can write the likelihood contribution of pool i as

$$\left\{\prod_{k=1}^{M_i-1}[S(T_{ik}; \bar{v}_{ik}, \theta) - S(T_{ik+1}; \bar{v}_{ik+1}, \theta)]^{P_{ik}}\right\} \times$$

$$S(T_{i,M_i}; \bar{v}_{i,M_i}, \theta)^{R_{M_i}}$$

where $\bar{v}_{ik}$ are the historical covariates for pool i until time $T_{ik}$. The likelihood function, in terms of contributions from each pool rather than from each individual mortgage, now becomes:

$$\left\{\prod_{k=1}^{M_i-1}\left[\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau, \theta, v(\tau)]d\tau\right) - \exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau, \theta, v(\tau)]d\tau\right)\right]^{P_{ik}}\right\} \times$$

$$\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau, \theta, v(\tau)]d\tau\right)^{R_{M_i}} \quad (4)$$

This is still a complicated functional form, equivalent to Equation (3). Moreover, there are still integrations of the time-dependent hazard functions required over long periods. Appendix A shows that Equation (4) can be further simplified to:

$$\prod_{k=1}^{M_i-1}\left\{\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau, \theta, v(\tau)]d\tau\right)^{R_{i,k+1}} \times \left[1 - \exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau, \theta, v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

Note that the likelihood function here includes only integrations over the short sampling intervals $(T_{i1}, T_{i2}]$, $(T_{i2}, T_{i3}]$, …, $(T_{iM_i-1}, T_{iM_i}]$ rather than over the interval from the time of mortgage origination to the time of prepayments as in Equations (3) and (4). This form of the likelihood function is computationally much more tractable.

For a sample of N pools, the total simplified log-likelihood is:

$$L(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{M_i-1}\left\{-R_{i,k+1}\left(\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau, \theta, v(\tau)]d\tau\right) + P_{ik}\ln\left[1 - \exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau, \theta, v(\tau)]d\tau\right)\right]\right\} \quad (5)$$

Our sampling interval is one month, because prepayment data are available only on a monthly basis. Since the integration interval is fairly short under the specification in Equation (5), we assume the covariates within this interval are constant. With this assumption we further simplify the log-likelihood function (5) to:

$$l(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{M_i-1}\left\{-R_{i,k+1}(T_{i,k+1} - T_{ik})\pi[T_{i,k+1}, \theta, v(T_{i,k+1})] + P_{ik}\ln\left[1 - \exp(-(T_{i,k+1} - T_{ik})\pi[T_{i,k+1}, \theta, v(T_{i,k+1})])\right]\right\} \quad (6)$$

Let $u_{ik} = \exp(-(T_{i,k+1} - T_{ik})\pi[T_{i,k+1}, \theta, v(T_{i,k+1})])$ and

$$y_{ik} = \frac{P_{ik}}{P_{ik} + R_{i,k+1}} = \frac{P_{ik}}{R_{i,k}}$$

This likelihood function (6) is identical to the case where $y_{ik}$ is the observed "success" rate following a binomial distribution with underlying probability of "success" $u_{ik}$, and $y_{ik}$ is independently distributed, conditional on $u_{ik}$. Therefore, estimating the prepayment survival model using the likelihood function (6) is equivalent to estimating the binomial GAM:

$$y_{ik} \sim \frac{1}{R_{i,k}} \text{binomial}(N = R_{i,k}, p = u_{ik})$$

$$u_{ik} = \exp\Big\{-(T_{i,k+1} - T_{ik}) \exp[\pi_0 + f_0[\text{age}_i(T_{i,k+1})] + f_1\left(\frac{c_i}{L(T_{i,k+1} - 0.25)}\right) + f_2[\overline{BO}_i(T_{i,k+1})] + f_3(\tilde{B}O_i) + \beta D(t)]\Big\}$$

The link function is $u_{ik} = \exp\{-(T_{i,k+1} - T_{ik}) \times \exp(\eta_{ik})\}$, which defines the relationship between $u_{ik}$ and the log hazard rate for pool i in month k, given by:

$$\eta_{ik} = \pi_0 + f_0[\text{age}_i(T_{i,k+1})] + f_1\left(\frac{c_i}{L(T_{i,k+1} - 0.25)}\right) + f_2\left[\overline{BO}_i(T_{i,k+1})\right] + f_3(\tilde{B}O_i) + \beta D(t)$$

We use a non-parametric local scoring method to estimate the four non-parametric components $f_0$, $f_1$, $f_2$, and $f_3$ simultaneously. The local scoring algorithm iteratively fits weighted non-parametric components by backfitting. The backfitting algorithm iteratively smooths partial residuals using a scatterplot smoother such as kernel regression. (Local scoring is discussed in Hastie and Tibshirani [1986].)

In our application, we modify the link function and error function as well as the reweighting procedure to suit our prepayment data. The scatterplot smoother we choose is a kernel regression. We use a Gaussian kernel and choose the bandwidths for each of the four functions via cross-validation (see Appendix B for the details of the local scoring algorithm and the scatterplot smoother algorithm).[5]

## III. DATA AND RESULTS

### Data

We first obtain the identification numbers for GNMA pools backed by thirty-year level-payment mortgages on single-family homes using BridgeStation's bond search engine. Then we retrieve prepayment history for these pools from Bloomberg. The criteria used in selecting pools are 1) the origination balances of the pools should be at least $15 million; and 2) the issuance dates should be spread over time from 1970 to 1998. Among these pools, we select a sample with wide variations in coupon rates.

We are able to obtain original size, date of origination, coupon rate, prepayment history, and average age at origination for thirty-nine of these pools. The prepayment histories for these pools are obtained over the June 1971-July 1998 sample period.

For each pool, we compute the mortgage origination date by subtracting the average mortgage age at pool origination date from the pool origination date. The mortgages in our sample were issued between November 1968 and March 1996. Therefore, we have a fairly long time series of prepayment histories and prepayment data for mortgages at different ages.[6] The mortgage coupons ranged from 6% to 16.5%.

Exhibit 1 presents the time series of mortgage coupons. The high-coupon mortgages originated in the late eighties and early seventies, and some of these mortgages were not fully prepaid even as late as 1998 when the mortgage rate was around 7%.

We use thirty-year constant-maturity yield data to proxy for refinancing rates (obtained from the St. Louis Federal Reserve Bank database). Following Schwartz and Torous, we use the average of the thirty-year constant-maturity rate in the three months preceding the observation month as proxies for the refinancing rate rather than the rate at the time of observation.

A monthly time series of the constant-maturity yield on the thirty-year Treasury bond from February 1977 through July 1998 was obtained from the Federal Reserve Bank data source. Thirty-year constant-maturity yields between June 1971 and January 1977 were not available. We therefore estimate them by fitting a linear regression of thirty-year constant-maturity yields on ten-year constant-maturity yields, which were available before January 1971.

### Results

We estimate the model in two stages. In the first stage, we ignore the effect of burnout and estimate the model:

$$\ln \pi_i(t) = \pi_{01} + f_{01}[age_i(t)] + f_{11}\left(\frac{c_i}{L(t-0.25)}\right) + \beta_1[summer(t)]$$

where L is the thirty-year constant-maturity yield, and summer is a dummy variable for summer.

We use the estimates from the Stage 1 model to compute expected burnout $\overline{BO}_i(t)$ for each pool. The unexpected burnout $\widetilde{BO}_i(t)$ is the difference between the realized and the predicted burnout. In Stage 2 we estimate the complete model:

$$\ln \pi_i(t) = \pi_{02} + f_{02}[age_i(t)] + f_{12}\left(\frac{c_i}{L(t-0.25)}\right) + f_2\left[\overline{BO}_i(t)\right] + f_3[\widetilde{BO}_i(t)] + \beta_2[summer(t)]$$

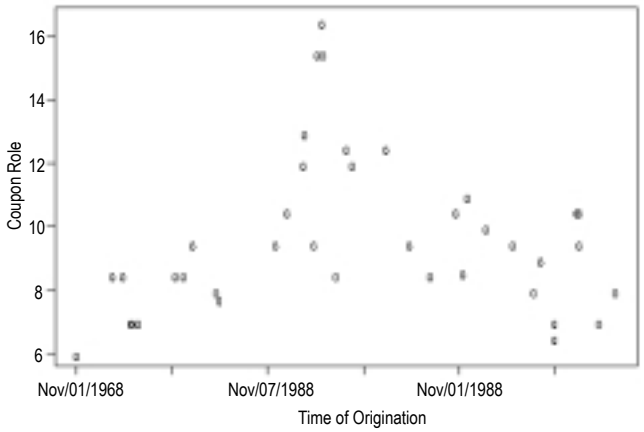
By employing the local scoring technique described in Appendix B, we estimate the parameters $\pi$ and $\beta$ and the non-parametric functions for Stage 1 and Stage 2 models. Exhibit 2 presents the estimates of the intercept and $\beta_1$ (summer dummy slope coefficient) for the Stage 1 model. Exhibit 3 presents these parameter estimates for the Stage 2 model.

The point estimate of $\beta_2$ is 0.12, which indicates that the hazard rate in summer months is 12.7% (= $e^{0.12}$) higher than in non-summer months. Since the residuals are likely to be correlated cross-sectionally (possibly due to omitted factors) we use Efron [1982] jackknifing techniques to determine confidence intervals (bands) and standard errors.

These standard errors are estimated as follows. The length of the sample period is 326 months. We leave out one month at a time, and estimate the model 326 times. We use the distribution of the estimates from these 326 experiments, adjusted suitably for degrees of freedom as proposed by Efron, to arrive at the confidence intervals (bands). The lower limit of the 95% two-sided confidence interval is 0.039, indicating that $\beta_2$ is reliably larger than zero, consistent with our discussion earlier.

Exhibit 4 plots the baseline hazard function. The hazard rate increases almost linearly up to around forty months, and then remains fairly stable up to month 270. From month 270 to month 360, the hazard rate increases rapidly. Because of the intercept term and the log specification, this function cannot be directly translated into monthly prepayment rates in familiar CPR units. To present the baseline hazard function in a form that is more

## EXHIBIT 1
**Distribution of the Pools in the Sample**




## EXHIBIT 2
**Stage 1 Model Parameter Estimates**

| Parameters | Estimate |
|---|---|
| $\pi_{01}$ | –7.78735 |
| $\beta_1$ | 0.09746 |

## EXHIBIT 3
**Stage 2 Model Parameter Estimates**

| Parameter | Estimate | Standard Error | 95% Confidence Interval |
|---|---|---|---|
| $\pi_{02}$ | –9.632 | 0.311 | –10.252<br>–8.931 |
| $\beta_2$ | 0.120 | 0.051 | 0.039<br>0.234 |

Standard errors and 95% confidence intervals are obtained using Efron's [1982] jackknife procedure.

intuitive, Exhibit 5 plots the estimated prepayment rate (in units of CPR) for a mortgage with 8.5% coupon and L = 8.0%. To isolate the relation between age and prepayments, this graph assumes that the interest rates remain unchanged over the life of the mortgage and that the unexpected burnout is zero.

The sawtooth pattern in Exhibit 5 is due to the seasonal summer spike in prepayments. The basic shape of the baseline hazard function, though, is similar to the PSA experience from age 0 to about 270 months. There are some minor deviations over this age range, and our model predicts higher starting prepayment rates for infant pools. The PSA model does not capture the rapid rise in baseline prepayment rates beyond 270 months, when the prepayment rates can be four times as high as PSA experience.

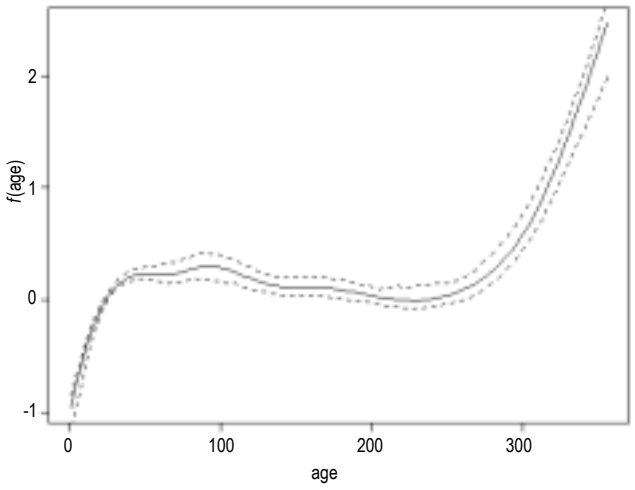
This shape of the hazard function makes intuitive sense. To understand the intuition behind this result, keep in mind that the baseline hazard rate characterizes prepayments that are unrelated to changes in interest rates. Early in the life of a mortgage, prepayments are fairly low since homeowners who take out new mortgages typically do not plan to move soon. Later on, more and more homeowners experience marriages, divorces, or job changes, and then sell their homes and prepay. The leveling off of the prepayment rates after about thirty months suggests that the extent to which homeowners can anticipate housing related changes in their personal situation thirty months down the line is about the same as what they can anticipate 270 months from origination.

The rapid increase in hazard rates beyond month 270 occurs because most homeowners tend to prepay their mortgages after building sufficient equity in their homes. By this time, the principal outstanding in level payment thirty-year mortgages is relatively low, and the homeowners, who also age, tend to have larger savings and simply extinguish their mortgage loans.[7] The rich characterization of the baseline hazard function that we obtain using nonparametric techniques would be very difficult to obtain using any of the conventional parametric specifications.

Exhibit 5 also presents the S&T log-logistic baseline hazard function for the mortgage pool. The estimates for this figure are obtained with our data. The log-logistic function captures the early rise in prepayments. It peaks at thirty-six months, and declines rapidly from that point on. The log-logistic function is unimodal and it is not flexible enough to capture the rapid rise in prepayment rates of relatively mature mortgages. Also, the estimated log-logistic declines steeply after peaking at forty months, and the prepayment rates implied by this function are significantly below what we estimate here and also below the PSA experience for older mortgage pools.

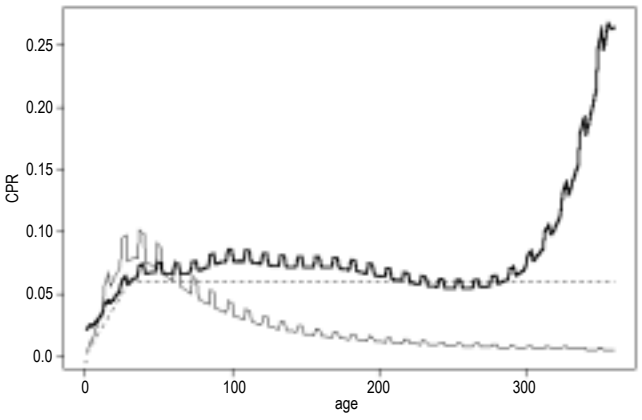
Exhibit 6 presents the relation between the hazard rate and the long-term interest rate (the benefits of refinancing). The hazard rate increases linearly up to a 1.5

## EXHIBIT 4
**The Baseline Hazard Function**




Solid line is estimated baseline hazard function. Dashed lines denote 95% confidence band.

## EXHIBIT 5
**Baseline Prepayment Function in CPR**




Darker solid line shows estimated baseline function for our model. Broken line corresponds to 100% PSA experience. Thin solid line is baseline prepayment rate for this pool based on Schwartz and Torous [1989] prepayment model.

ratio of mortgage coupon to long rate. Beyond this point, there is no further increase. This result is also fairly intuitive. When the level of refinancing rate is close to the mortgage coupon rates, mortgagors prepay if the benefits from prepayment exceed their transaction costs. As the benefit increases, a larger proportion of the homeowners in the pool end up prepaying their mortgages.

Beyond a certain level, the refinancing benefits are so great that simple financial transaction costs are not sufficient to explain why some mortgagors are still in the pool. The mortgagors in the pool at this point are probably there because of bad credit situations or because they are simply not aware of low mortgage rates. The prepayment probabilities of these mortgagors would likely depend less on the extent of prepayment benefits and more on how likely they are to improve their credit quality or how likely they are to become aware of prevailing refinancing rate. These factors explain why hazard rates level off at higher levels of c/L.

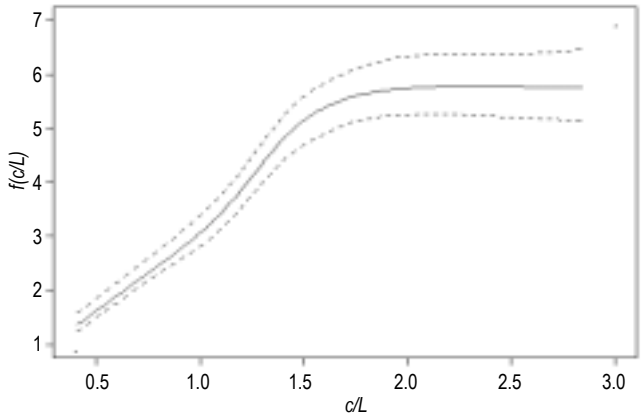
The confidence band plotted in Exhibit 6 indicates this function is more accurately estimated for lower levels of c/L than for higher levels. This is because we have very few pools that are exposed to c/L ratios higher than 1.5. Only the high-coupon pools that originated in late seventies and early eighties had this experience.

Exhibit 7 plots the relation between the hazard rate and expected burnout. Except for very low levels of burnout, high burnout is associated with low hazard rate, as we expect. If the pool experiences a low interest rate environment, we would expect the highly sensitive prepayers to leave the pool and those remaining in the pool are slow to prepay. There is no obvious reason why the hazard rate increases at low levels of expected burnout, but as can be seen from Exhibit 7 the confidence interval is fairly wide in this region, and it is difficult to draw strong conclusions from the shape of the function here.

Exhibit 8 plots the relation between the hazard rate and unexpected burnout. The unexpected burnout captures pool-specific variations in prepayment tendencies. There is a monotonically increasing relation between unexpected burnout and the hazard rate, which implies that unexpectedly fast prepayment pools continue to be fast prepayment pools for the rest of their lives. As indicated by the narrow confidence interval bands, we are able to estimate this function fairly accurately.

Exhibit 9 evaluates the goodness of fit of this model. The average error (the difference between fitted and actual prepayment rate expressed in SMM) is close to zero, indicating that the estimates are unbiased. The median is also close to zero, indicating that we are as likely to find positive errors and negative errors. The interquartile range is 0.44%. We compute the $R^2$ of this model as the ratio of the variance of the fitted prepayment rates over variance of the actual prepayments. The $R^2$ for our model is 69.4%.

**EXHIBIT 6**
**Relation Between Hazard Rate and Long-Term Interest Rate**

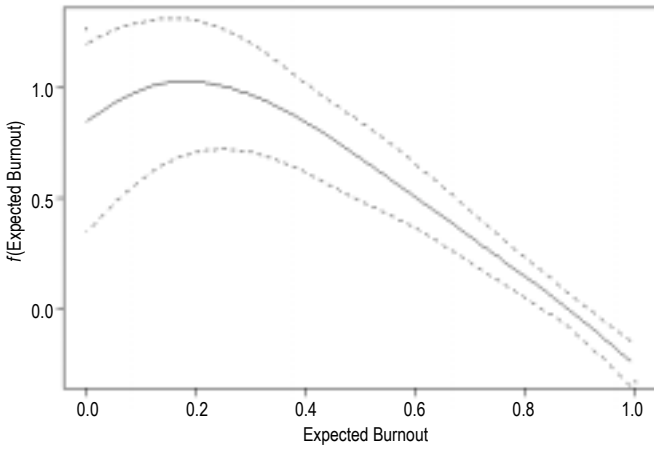



Solid line is estimated function of long-term interest rate. Dashed lines denote 95% confidence band.

To put the goodness of fit of our model into perspective and to quantify the benefits of our approach over the parametric methods in the literature, we compare its performance with the Schwartz and Torous [1989] model.[8] The row titled S&T model with incomplete likelihood function presents the results when their model is fitted to our data ignoring the time-dependence of the covariates. The interquartile range is 1.6% for this model and the $R^2$ is only 3.7%. This result indicates that our model represents a significant improvement over the S&T model.

To evaluate the extent to which the benefits of our methods are due to the use of a complete likelihood function and the extent to which the improvements are due to our non-parametric approach, we also evaluate the performance of the S&T model using the reformulated complete likelihood function. The $R^2$ of the S&T model increases to 36.8%, and the interquartile range falls to 0.857%.[9] Using the incomplete likelihood function for estimation clearly results in significant loss of information, which adversely affects estimation efficiency. These results indicate that the benefits of using our model are due both to the use of a complete likelihood function and to

## EXHIBIT 7
**Relation Between Hazard Rate and Expected Burnout**




Solid line is estimated function of expected burnout rate. Dashed lines denote 95% confidence band.

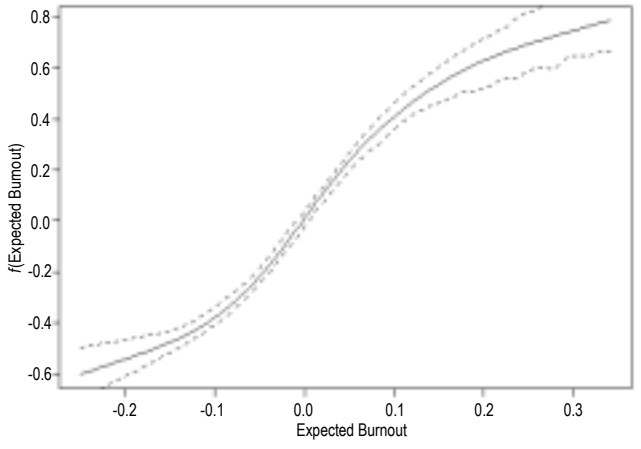
the use of non-parametric methods that allow flexible functional relations between the covariates and hazard rates and path-dependent covariates.

Exhibit 10 examines the goodness of fit of our model for pools with different coupon rates. This table is based on the model estimated with the entire sample; the model is not fitted separately for each subsample. The mean is close to zero within each subsample of mortgage pools, indicating that there are no biases in our estimates related to mortgage coupons. The interquartile range for the low coupon pools is 0.365% and for the high-coupon pool 0.577%. There is a monotonically increasing relation between the interquartile range and coupon rate. This is to be expected, since the prepayment rates for high-coupon pools typically tend to be larger and more variable.

Exhibit 11 presents the performance of the model over three subperiods. Since we have many more observations in the eighties and nineties, the prepayment rates are better predicted in those periods. The higher interquartile ranges for the seventies and the nineties occur because the interest rates in these periods were low, and as a result prepayments were more frequent and more variable.

To provide an overall perspective on the goodness of fit of our model, Exhibit 12 plots the prepayments averaged across all pools and the predicted prepayments based on our model. Prepayments are low in the late seventies and early eighties because of high interest rates. With declining interest rates in the early nineties, prepayment picks up rapidly. The next round of low interest rates in the mid- to late-nineties did not witness the same level of high prepayments as in the early nineties because of the high burnout rates. As Exhibit 12 illustrates, prepayment rates estimated by our model closely track the observed pattern.

## EXHIBIT 8
**Relation Between Hazard Rate and Unexpected Burnout**




Solid line is estimated function of unexpected burnout rate. Dashed lines denote 95% confidence band.

## IV. PREPAYMENT AND PRICES OF MORTGAGE-BACKED SECURITIES

To examine the relation between various factors that affect prepayments and MBS prices using our prepayment model, we use the Chen and Scott [1992] version of the two-factor Cox-Ingersoll-Ross interest rate model in pricing analysis. Under this model, the interest rate process is given by:

$$i_t = y_{1t} + y_{2t}$$

and

$$\begin{cases} dy_{1t} = \kappa_1(\theta_1 - y_{1t})dt + \sigma_1\sqrt{y_{1t}}dB_{1t} \\ dy_{2t} = \kappa_2(\theta_2 - y_{2t})dt + \sigma_2\sqrt{y_{2t}}dB_{2t} \end{cases}$$

$$dB_{1t}dB_{2t} = 0$$

## EXHIBIT 9
**Prepayment Prediction Errors and Model Comparison**

| Model | 25th Percentile | Median | 75th Percentile | Interquartile Range | Mean | $R^2$ |
|---|---|---|---|---|---|---|
| Our Model | –0.00259 | –0.00069 | 0.00186 | 0.00444 | –2.987e–006 | 0.694 |
| S&T Model with Incomplete L.F. | –0.01155 | –0.00319 | 0.00467 | 0.01622 | –2.379e–003 | 0.037 |
| S&T Model with Complete L.F. | –0.00565 | –0.00130 | 0.00292 | 0.00857 | –5.973e–005 | 0.368 |

## EXHIBIT 10
**Prediction Error for Pools with Different Coupons**

| Coupon Range | Number of Pools | 25th Percentile | Median | 75th Percentile | Interquartile Range | Mean |
|---|---|---|---|---|---|---|
| ≤ 7 | 7 | –0.00171 | –0.00019 | 0.00194 | 0.00365 | 0.00032 |
| > 7, ≤ 9 | 12 | –0.00257 | –0.00063 | 0.00176 | 0.00433 | –0.00034 |
| > 9, < 12 | 12 | –0.00299 | –0.00108 | 0.00168 | 0.00467 | 0.00044 |
| ≥ 12 | 8 | –0.00348 | –0.00095 | 0.00229 | 0.00577 | –0.00044 |

## EXHIBIT 11
**Prediction Error Across Subperiods**

| Subperiod | Number of Observations | 25th Percentile | Median | 75th Percentile | Interquartile Range | Mean |
|---|---|---|---|---|---|---|
| 1971–1979 | 635 | –0.00135 | 0.00105 | 0.00346 | 0.00481 | 0.00132 |
| 1980–1989 | 2,531 | –0.00241 | –0.00101 | 0.00100 | 0.00341 | –0.00078 |
| 1990–1998 | 3,473 | –0.00299 | –0.00051 | 0.00273 | 0.00572 | 0.00049 |

where $i_t$ is the instantaneous rate at time t, driven by two uncorrelated stochastic factors $y_{1t}$ and $y_{2t}$. The two factors revert to their respective long-run means $\theta_i$ at speed $\kappa_i$. This process allows for a rich variety of interest rate dynamics and a wide array of term structures.[10]

For example, when $\kappa_1$ and $\kappa_2$ are small, the yield curves will be flat and they shift in a parallel manner. When $\theta_1$ and $\kappa_1$ are large and $\sigma_2$ is small, there will be little change in the long end of the yield curve, but the short end will be volatile. When the mean reversion factor $\kappa_2$ is small, $y_{2t}$ behaves like a random walk, and it plays the dominant role in the determination of long-term interest rates.

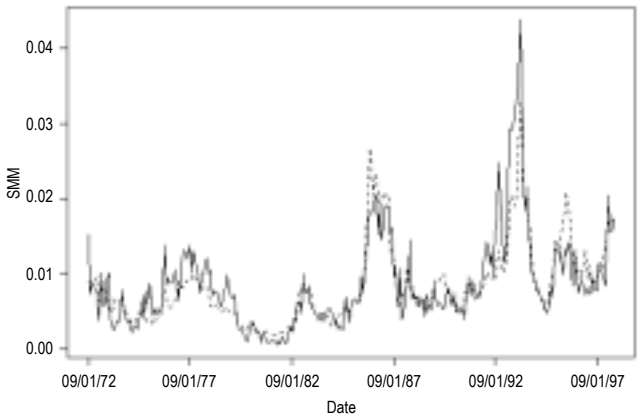
Chen and Scott show that the time t yield on a zero-coupon bond that matures at time s, denoted by Y(t, s) under this two-factor model, is:

$$Y(t, s) = -\frac{B_1(t, s)}{s - t} y_{1t} - \frac{B_2(t, s)}{s - t} y_{2t} + \frac{\ln A_1(t, s)}{s - t} + \frac{\ln A_2(t, s)}{s - t}$$

where

$$A_i(t, s) = \left[\frac{2\gamma_i \exp[(1/2)(\kappa_i + \lambda_i + \gamma_i)(s - t)]}{(\kappa_i + \lambda_i + \gamma_i)(\exp[\gamma_i(s - t)] - 1) + 2\gamma_i}\right]^{2\kappa_i\theta_i/\sigma_i^2}$$

## EXHIBIT 12
**Estimated and Observed Prepayment Rates**




Solid line is average observed prepayment rates. Dashed lines is estimated prepayment rate.



and

$$B_i(t, s) = \frac{2(\exp[\gamma_i(s-t)]-1)}{(\kappa_i+\lambda_i+\gamma_i)(\exp[\gamma_i(s-t)]-1)+2\gamma_i}$$

where $\lambda_i$ is the proportional risk premium for the i-th factor.

To price mortgage-backed securities, we generate 250 simulated time series of interest rates under the risk-neutral process for the i-th factor given by:

$$dy_{it} = [\kappa_i(\theta_i - y_{it}) - \lambda_i y_{it}]dt + \sigma_i\sqrt{y_{it}}\,dB_{it}$$

We use the process parameters estimated by Chen and Scott in our pricing analysis. Specifically, the parameters we use are $\kappa_1 = 1.8341$, $\theta_1 = 0.05148$, $\sigma_1 = 0.1543$, $\lambda_1 = -0.1253$, $\kappa_2 = 0.005212$, $\theta_2 = 0.03083$, $\sigma_2 = 0.06689$, and $\lambda_2 = 0$.[11]

### MBS Prices and the Seasoning Effect

An analysis of how MBS prices vary through the seasoning process will illustrate the importance of correctly identifying the baseline hazard function in order to obtain good MBS prices.

We simulate the thirty-year interest rate path under the risk-neutral interest rate dynamics using the parameters listed above. The starting factor values are $y_1 = 0.05525$ and $y_2 = 0.03083$, which are their unconditional mean values under our process parameters. We then determine the prices of two MBS, one with a mortgage coupon of 11%, and the other with a mortgage coupon of 7%. We choose these pools so that one pool has a coupon lower than the refinancing rate at the start of the simulation (and also expected refinancing rates at different points in the life of the mortgage), and the other pool has a higher coupon. We set the unexpected burnout to zero at all ages.

Exhibit 13A presents the prices (per dollar of face value) of the 11% MBS at different ages based on 250 simulations. The MBS price is the highest at the beginning, because this pool coupon is higher than the refinancing rate, and the prepayment rate is the slowest at this point (see Exhibit 4). Progressively, as the prepayment rate rises, the MBS price declines toward its face value. The decline in the MBS price slows down after about month 30, and at this time we can also see the baseline hazard rate leveling off in Exhibit 4. Finally, the MBS price reaches its face value at maturity.

Exhibit 13B presents the price-age curve for the 7% MBS. Since the coupon rate is lower than the current yield curve, the MBS trades at a discount. The price of the MBS rises over time in almost a mirror image of the drop in prices for the 11% MBS. For instance, in the early years and in the late years there is a faster rise in prices driven by rapid increases in prepayment rates.

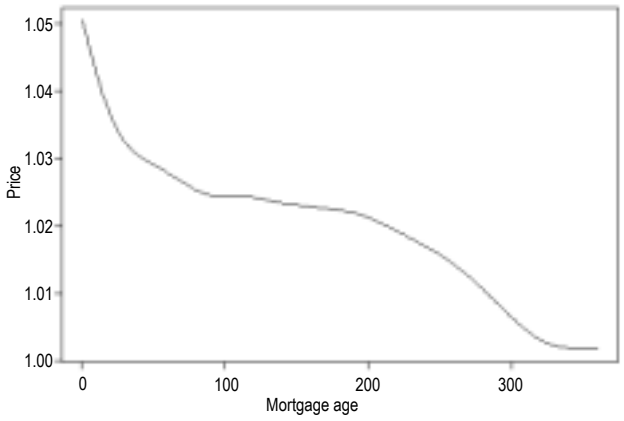
In general, we can expect the price of any premium bond to fall over time and the price of any discount bond to rise over time. What is interesting in the case of the MBS is that the rate of increase or decrease in prices with age is tightly driven by the shape of the baseline hazard function. Discount MBS prices rise faster when the prepayment rates are expected to rise rapidly; the converse is true for premiums.

### MBS Prices and the Burnout Effect

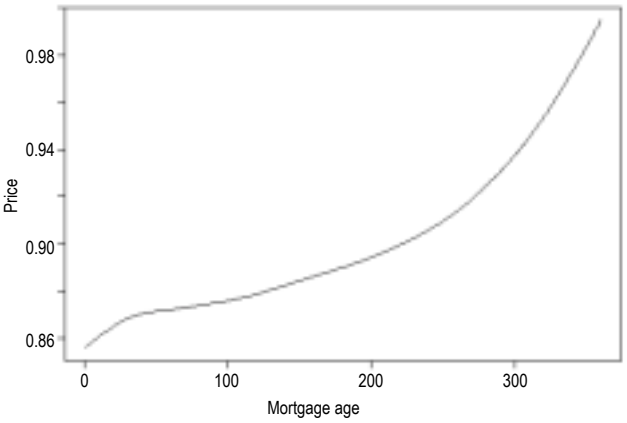
To examine the effect of burnout on MBS prices, we follow a similar simulation procedure, except that now we keep the age of the pool constant at forty-eight months but vary the refinancing rates at the start of the simulations.

We first consider the prices of two pools of 11% MBS at different levels of interest rates. Both Pools A and

## EXHIBIT 13A
**Price as a Function of Age — 11% MBS**



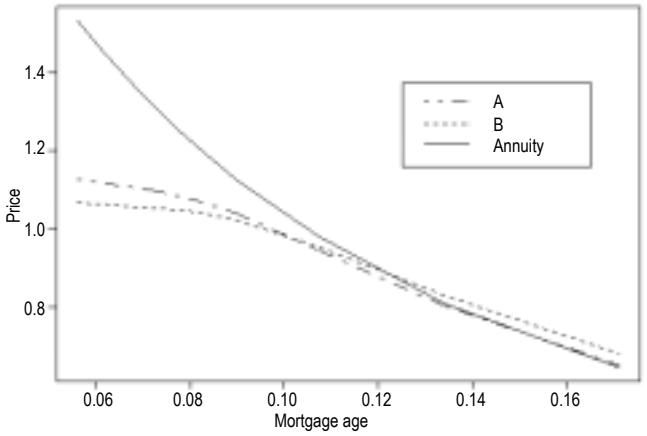

## EXHIBIT 13B
**Price as a Function of Age — 7% MBS**




B have a total burnout of 0.40. Pool A has an expected burnout of 0.55 and an unexpected burnout of –0.15, while Pool B has an expected burnout of 0.25 and an unexpected burnout of 0.15. Recall that higher expected burnout is associated with slower future hazard rates, as is lower unexpected burnout. Therefore, with these burnout assumptions, Pool B is expected to prepay faster in the future than Pool A.

Exhibit 14 presents the prices of these MBS at different levels of yields on annuities that trade at face value. For comparison, it also presents the price of a $1 face value 11% fully amortizing bond (or an annuity) with no prepayment option.

## EXHIBIT 14
**MBS Price and Pool Burnouts**




A: Expected burnout 0.55, unexpected –0.15.
B: Expected burnout 0.25, unexpected 0.15.

At low levels of interest rates, the prepayment option is highly valuable, and both pools trade at substantial discounts to the annuity. In fact, at yields below about 9% the MBS exhibit negative convexity because of the prepayment option. At very high interest rates, however, the MBS trade at prices slightly higher than straight annuities, despite this option. This is because the option effect is offset by the tendencies of some mortgagors to prepay even when rates are high. Holders of MBS benefit from prepayments in high interest rate environments, and hence they bid the prices of MBS above that of comparable annuities.

The prices of Pool A and Pool B differ at virtually all interest rate levels. At lower interest rate levels, Pool A is priced higher than Pool B. This is because when the current rate is lower than the pool rate, investors prefer slower prepayments. At higher current rates, investors prefer faster prepayments, and this is reflected by the fact that at these rates Pool B is priced higher than Pool A.

The results here highlight the importance of separately identifying the components of burnouts in order for accurate pricing of mortgage-backed securities.

## V. CONCLUSION

A mortgage prepayment model is a critical component of any valuation model for mortgage-backed

securities and their derivatives. The prepayment decisions of mortgagors depend on a number of factors such as the age of the mortgage and incentives for refinancing. Although researchers are well aware of the factors that influence prepayment decisions, there is no theory that predicts the functional relations between these factors and prepayment rates, which thus need to be estimated empirically.

We develop and estimate a non-parametric prepayment model. There are several significant features that differentiate our model and estimation techniques from other models. First, we use a non-parametric estimation approach. This approach is adapted from the generalized additive model estimation technique developed by Hastie and Tibshirani [1986]. Our approach enables us to non-parametrically estimate the relation between prepayment rates and four factors that drive prepayment rates. Our approach generalizes the multivariate density estimation (MDE) approach, which is typically much more data-intensive and hence cannot be directly used in applications such as ours where there are more than two factors. In addition, our approach allows us to estimate a mixed parametric and non-parametric model.

The relationship between prepayment rates and the variables in our model is highly non-linear, and it is difficult to capture these relations with parametric functions. Prepayments increase rapidly with age until about month 40 and then level off until about month 270. The prepayment rate rises significantly from month 270 to month 360. The relationship between coupon rates and interest rates is also highly non-linear. Prepayments increase until the ratio of coupons to interest rate is 1.5, and then level off.

Our model decomposes the burnout effect into two components: expected burnout, which captures the effect of historical exposure of a pool to prepayment risks, and unexpected burnout, which captures heterogeneity across pools. We find that high expected burnouts lead to slower future prepayments but high unexpected burnouts are related to higher future prepayments. Overall, our model fits the data significantly better than models currently in use.

We also examine the relation between the factors that affect prepayments and the prices of mortgage-backed securities. There is a close relation between the estimated shape of the baseline hazard function and the theoretical prices of mortgage-backed securities with different ages. Decomposition of pool burnouts into expected and unexpected components also has important pricing effects.

# APPENDIX A

## Simplification of Equation (4)

$$\left\{\prod_{k=1}^{M_i-1}\left[\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)-\exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}\times$$

$$\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}=\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}\times$$

$$\prod_{k=1}^{M_i-1}\left\{\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)\left[1-\exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]\right\}^{P_{ik}}$$

$$=\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}\times$$

$$\prod_{k=1}^{M_i-1}\left\{\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{ik}-R_{i,k+1}}\left[1-\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

$$=\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}\prod_{k=1}^{M_i-1}\left\{\frac{\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{ik}}}{\exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}}\times\right.$$

$$\left.\frac{\exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}}{\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}}\left[1-\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

$$=\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}\prod_{k=1}^{M_i-1}\left\{\frac{\exp\left(-\int_{D_i}^{T_{ik}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{ik}}}{\exp\left(-\int_{D_i}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}}\times\right.$$

$$\left.\left[\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{R_{i,k+1}}\left[1-\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

$$=\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{M_i}}\frac{\exp\left(-\int_{D_i}^{T_{j1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i1}}}{\exp\left(-\int_{D_i}^{T_{i,M_i}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,M_i}}}\times$$

$$\prod_{k=1}^{M_i-1}\left\{\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}\left[1-\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

$$=\prod_{k=1}^{M_i-1}\left\{\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)^{R_{i,k+1}}\left[1-\exp\left(-\int_{T_{ik}}^{T_{i,k+1}}\pi[\tau,\theta,v(\tau)]d\tau\right)\right]^{P_{ik}}\right\}$$

The last step follows because $T_{i1} = D_i$.

# APPENDIX B

## Algorithms Used in Non-Parametric Generalized Additive Model

### Local Scoring Procedure for Prepayment Model

Step 0: Initialize

$$\pi_0^{(0)} = g\left(\frac{1}{\sum_i M_i}\sum_{i,k} y_{ik}\right);\ f_0^{(0)}, f_1^{(0)}, f_2^{(0)}, f_3^{(0)}, \beta^{(0)} = 0$$

where g, defined as

$$\eta_{ik} = g(u_{ik}; T_{i,k+1} - T_{i,k}) = \log\left(-\frac{\log u_{ik}}{T_{i,k+1} - T_{i,k}}\right)$$

is the inverse function of the link function.

At Step j: Construct adjusted dependent variables:

$$z_{ik}^{(j)} = \eta_{ik}^{(j-1)} + (y_{ik} - u_{ik}^{(j-1)})g'(u_{ik}^{(j-1)})$$

where

$$\eta_{ik}^{(j-1)} = \pi_0^{(j-1)} + \sum_{q=0}^{3} f_q^{(j-1)}(X_{q,ik}) + \beta^{(j-1)}\text{Summer}_{ik}$$

and

$$u_{ik}^{(j-1)} = g^{-1}(\eta_{ik}^{(j-1)})$$

Here $X_{q,ik}$ stands for the q-th covariate for pool i at time k.

Construct weights

$$w_{ik}^{(j)} = \left[g'(u_{ik}^{(j-1)})\right]^{-2}\left(\frac{u_{ik}^{(j-1)}(1-u_{ik}^{(j-1)})}{R_{i,k}}\right)^{-1}$$

Let $f_q^{(j)} = f_q^{(j-1)}$, q = 0, 1, 2, 3. Estimate $f_q^{(j)}$, q = 0, 1, 2, 3, by iteratively running scatterplot smoother:

$$f_q^{(j)} = \text{Smoother}\left(z_{ik}^{(j)} - \pi_0^{(j-1)} - \sum_{q'\neq q}^{3} f_{q'}^{(j)}(X_{q',ik}) - \beta^{(j-1)}\text{Summer}_{i,k}\middle| X_{q,ik},\ \text{weight} = w_{ik}^{(j)}\right)$$

with q = 0, 1, 2, 3, 0, 1, 2, 3, …, until convergence.

Estimate $\pi_0^{(j)}$ and $\beta^{(j)}$ by fitting weighted least square estimator (WLSE):

$$\{\pi_0^{(j)}, \beta^{(j)}\} = \text{WLSE}\left(z_{ik}^{(j)} - \sum_{q'=0}^{3} f_{q'}^{(j)}(X_{q',ik})\middle|\{1, \text{Summer}_{i,k}\}, \text{weight} = w_{ik}^{(j)}\right)$$

Stop if the differences between $f_q^{(j)}$ and $f_q^{(j-1)}$ and the difference between $(\pi_0^{(j)}, \beta^{(j)})$ and $(\pi_0^{(j-1)}, \beta^{(j-1)})$ are small enough. Otherwise, j = j + 1; return to step j.

### Weighted Scatterplot Smoothing Procedure

The weighted scatterplot smoother can be chosen from a variety of candidates including cubic-spline, B-spline, local regression, or kernel regression. We use Gaussian kernel regression here.

For a sample of $(y_i, x_i)$, i = 1…n, and weights $w_i$, a kernel regression can give a smoothed function y = f(x) with

$$f(x; v) = \frac{\sum_{i=1}^{n} y_i w_i K(x - x_i; v)}{\sum_{i=1}^{n} w_i K(x - x_i; v)}$$

where

$$K(x - x_i; v) = \frac{1}{\sqrt{2\pi}v} e^{-\frac{(x-x_i)^2}{2v^2}}$$

The bandwidth, v, is a positive number chosen by solving this cross-validation problem suggested by Hardle [1990]:

$$\min_v \sum_{i=1}^{n} (f_{(y_i,x_i)\ \text{excluded}}(x_i; v) - y_i)^2$$

## ENDNOTES

The authors thank David Ling and seminar participants at Fannie Mae, the University of Florida at Gainesville, University of Illinois at Urbana-Champaign, University of Maryland, Purdue University, and University of Utah for helpful comments.

[1]Articles on rational prepayment with option pricing approaches include Dunn and McConnell [1981], Timmis [1985], Johnston and Van Drunen [1988], and Stanton [1995].

[2]To further understand the "dimensionality curse" of non-parametric estimation techniques, suppose one needs to allow for four independent variables, which is not uncommon in our context. If one divides the domain of each variable into only forty equally spaced pieces, this would yield a total of $40^4$ = 2.56 million neighborhoods. So one would need at least 2.56 million observations to ensure an average of just one data point per neighborhood in the MDE approach.

[3]Recall that the mortgages that underlie GNMA MBS are fully amortizing loans, so even in the absence of prepayments the outstanding principal will decline over the life of a mortgage.

[4]For a general introduction to survival analysis, see Kalbfleisch and Prentice [1980].

[5]See Hardle [1990] for details of kernel regressions and cross-validation.

[6]During the early years of our sample period, in some cases, data were not available for some pools in some months. Because our estimation technique can handle censored data, we do not exclude these pools but treat these observations as two-side-censored data.

[7]For example, for a 270-month-old 8.0% mortgage, only 49.5% of the original principal is still outstanding.

[8]We compare our model with S&T because that model is currently considered the benchmark in the literature.

[9]When we estimate the S&T model using the same data used in S&T, we also find that the interquartile range of the error terms is 61.9% lower when we use the full likelihood function.

[10]Longstaff and Schwartz [1992] propose a similar two-factor model where one factor is identified with the instantaneous rate and the other factor is identified with stochastic volatility.

[11]Chen and Scott estimate $\lambda_2 = -0.0665$, but we set $\lambda_2 = 0$ to avoid negative mean reversion for the second factor under the risk-neutral process.

## REFERENCES

Ait-Sahalia, Y. "Nonparametric Pricing of Interest Rate Derivatives Securities." *Econometrica*, 64 (1996), pp. 527-560.

Archer, W.R., and D.C. Ling. "Pricing Mortgage-Backed Securities: Integrating Optimal Call and Empirical Models of Prepayment." *Journal of the AREUEA*, 21 (1993), pp. 373-404.

Archer, W.R., D.C. Ling, and G. McGill. "The Effect of Income and Collateral Constraints on Residential Mortgage Terminations." *Regional Science and Urban Economics*, 26 (1996).

Boudoukh, J., R. Whitelaw, M. Richardson, and R. Stanton. "Pricing Mortgage-Backed Securities in a Multifactor Interest Rate Environment: A Multivariate Density Estimation Approach." *Review of Financial Studies*, 10 (1997), pp. 405-446.

Chen, R.-R., and L. Scott. "Pricing Interest Rate Options in a Two-Factor Cox-Ingersoll-Ross Model of the Term Structure." *Review of Financial Studies*, 5 (1992), pp. 613-636.

Dunn, K.B., and J.J. McConnell. "A Comparison of Alternative Models for Pricing GNMA Mortgage-Backed Securities." *Journal of Finance*, 36 (1981), pp. 471-483.

Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM, 1982.

Green, J., and J.B. Shoven. "The Effect of Interest Rates on Mortgage Prepayments." *Journal of Money, Credit and Banking*, 18 (1986), pp. 41-59.

Hardle, W. *Applied Nonparametric Regression*. Cambridge: Cambridge University Press, 1990.

Hastie, T., and R. Tibshirani. "Generalized Additive Models." *Statistical Sciences*, 1 (1986), pp. 297-318.

Hayre, L., and A. Rajan. "Anatomy of Prepayments: The Salomon Brothers Prepayment Model." Salomon Brothers, 1995.

Johnston, E., and L. Van Drunen. "Pricing Mortgage Pools with Heterogeneous Mortgages: Empirical Evidence." Working paper, University of Utah, 1988.

Kalbfleisch, J., and R. Prentice. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, 1980.

Longstaff, F., and E. Schwartz. "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model." *Journal of Finance*, 47 (1992), pp. 1259-1282.

Schwartz, E.S., and W.N. Torous. "Prepayment and the Valuation of Mortgage-Backed Securities." *Journal of Finance*, 44 (1989), pp. 375-392.

Stanton, R.H. "Rational Prepayment and the Valuation of Mortgage-Backed Securities." *Review of Financial Studies*, 8 (1995), pp. 677-708.

Timmis, G.C. "Valuation of GNMA Mortgage-Backed Securities with Transaction Costs, Heterogeneous Households and Endogenously Generated Prepayment Rates." Working paper, Carnegie Mellon University, 1985.