Comparing loan-level and pool-level mortgage portfolio analysis

Shirish Chinchalkar and Roger M. Stein

Moody's Research Labs

New York

Working Paper #2010-11-1

First draft: July 3, 2010

Current Draft: November 14, 2010

Abstract

With the help of newly-available tools, we show that using pool-level data rather than loan-level data for mortgage portfolio analysis can lead to substantially different conclusions about the credit risk of the portfolio. This finding is timely as there is an increased interest from market participants and regulators in improved credit risk models for this asset class. Further, recent advances in data accessibility as well as changes in regulatory reporting requirements have led to the increased availability of loan-level mortgage data, facilitating development of loan-level models to achieve higher accuracy in estimating portfolio credit risk. A number of theoretical results, some of which are recent, suggest that loan-level analysis may produce more reliable results than pool-level analysis. We verify this empirically. We present some evidence from a set of experiments that we performed using both approaches.

In addition to showing the impact of differences in analytic approach on mortgage portfolios, we use newly introduced software technology to assess impact on tranches of an RMBS transaction.

Our findings suggest that, when feasible, loan-level analysis produces more detailed default risk estimates in comparison to pool level analysis due to its more detailed representation of heterogeneous portfolios. This finding holds for a variety of realistic portfolio construction approaches. This finding persists even when we further subset a mortgage pool into dozens of sub-pools and correspondingly use dozens of customized summaries. These results have direct application in the area of retail mortgage credit risk management, RMBS valuation and investment selection.

1 Introduction[1]

Recent events in the financial markets have prompted market participants to seek more detailed approaches to estimating future losses on residential mortgage pools. Historically, mortgage portfolios were often analyzed at the *pool* level using summary statistics for characteristics of the underlying loan (e.g., LTV, credit score, loan type and geographic concentration)[2]. This approach contrasts with how evaluation is typically done, on an asset-by-asset basis, for corporate bond and loan portfolios at many institutions. For mortgages, it was often felt that, given the typically larges size of portfolios, aggregate-level analysis provided a reasonable approximation to the results of full analysis and aggregates could be applied uniformly across all mortgage portfolios, rather than only those for which loan-level data was available. Market participants are now increasingly interested in moving toward an asset-by-asset approach to evaluate residential mortgage pools at the loan-level as well.

This increased interest is motivated by concerns about: 1) the *heterogeneity* of underlying mortgages in a portfolio that may be masked by aggregate statistics and 2) the *interaction* of loan-level risk factors that may result in a higher (or lower) default risk for an individual mortgage and portfolio than that suggested by the individual loan characteristics in isolation. (This interaction is sometimes referred to as "layered risk" in the RMBS industry.)

Although the heterogeneity and interaction effects are difficult to capture with an aggregate-level model, the magnitude of the impact of these effects was not well documented and was often believed to be modest. Furthermore, in the case of RMBS transactions, it was often difficult to port the results of a loan-level analysis of a mortgage pool to the analytic tools used for evaluating the waterfalls (liabilities) of RMBS securities since no vendors provided this full integration natively. Thus, making the transition from loan-level analysis on the assets to pool-level analysis on the liabilities required users to make a number of assumptions, which could reduce the benefits of the loan-level analysis. For instance, many RMBS analysts use pool-level aggregate models when evaluating tranches (liabilities) of RMBS transactions, even when they have evaluated the assets at the loan-level.[3] (In some cases, institutions might use loan-level data within a waterfall tool to more realistically capture the *amortization schedule* of

[1] We wish to thank Grigoriy Enenberg for preparing the data used in this paper and Mihir Patel and Tamara Lubomirsky for assistance in running simulations on RMBS cashflow transactions using Moody's Wall Street Analytics waterfall platform. We also thank Navneet Agarwal, Ashish Das, Jacob Grotta, Andrew Kimball, Aziz Lookman and Mark Zandi for helpful comments on earlier drafts of this paper. We had useful econometrics discussions with Andrew Lo on aggregation and Members of Moody's Academic Advisory and Research Committee (MAARC) provided useful feedback on an earlier version of this research. All errors are of course our own.

[2] This is not uniformly the case. The rating agency Moody's Investors Service, for example, has performed loan-level analysis on mortgage pools since about 2002.

[3] Recent advances in modeling RMBS transactions have made it possible to perform loan-level analysis and simulation on the liabilities as well. For example, Moody's Analytics Structured Finance Workstation™ has recently introduced this functionality.

loans in the portfolio, but would still assume common aggregate *prepayment and default vectors* across all loans.)

The analysis of the differences between pool- and loan-level approaches is germane because the new results we present here suggest that the two methods do not generally produce similar results. To demonstrate the differences in outcomes that these approaches produce, we present results of a number of experiments we performed on identical mortgage portfolios using loan-level and aggregate portfolio representations. We performed aggregation at two levels, so readers can also compare how the aggregate measures perform as the level of aggregation changes. We use data drawn from a large cross section of prime RMBS mortgage pools and we examine the effects in two ways. We first explore the effects from an historical (*descriptive*) perspective and then examine the impact in a risk management (*forecasting*) setting, using models of mortgage behavior. These latter analyses are done using loan-level analytic tools to predict losses on pools of residential mortgage loans.

To anticipate the results we present later in the paper: we find that loan-level analysis produces more realistic behavior in the historical setting. Our analysis suggests that even for pools that are similar at the aggregate level, there are substantial and economically meaningful differences in credit quality, and that performing the credit analysis at the loan level captures many of these differences while analysis at the pool level may mask them. In forecasting settings the loan-level and pool-level results can also be quite different.

We also document the size of these aggregation effects. For example, when we compare the results of these two approaches. In our experiments, it was not uncommon to find relative differences in default rate behavior of 20%-40%. This is the case both when examining historical data and risk-management forecasts. Furthermore, it appears that these differences cannot be addressed through a general calibration since they arise from interactions of the various factors that drive credit risk, which may result in differences that are either positive or negative. These differences in analyses are particularly pronounced for portfolios with a greater degree of heterogeneity: those that contain systematic combinations of different levels of factors that are significant determinants of credit risk, such as LTV and FICO. These results suggest that substantially more assumptions must be made by analysts in choosing stress scenarios for aggregate analyses if they are performed.

We find that the dominance of loan-level analysis persists even when we further subset a pool into dozens of sub-pools for aggregation. A common form of aggregation used in industry involves converting cohorts of individual loans into a single "average" loan in order to predict the prepayment, default and severity for the mortgage pool. These "average" loans are termed *rep lines*, *stratifications* or *strats*.[4] For example, an analyst may create two rep lines for a mortgage pool: one for fixed-rate mortgages and one for adjustable rate mortgages to explicitly model differences in, say, the prepayment behavior of these two types of loans. While analysts may theoretically define an unlimited number of rep lines, in practice, most use only a handful.

[4] We adopt the convention of labeling these aggregations "rep lines" for the remainder of this paper.

We find evidence of pronounced differences in credit risk even when the number of sub-pools is much greater than often used in practice. For example, there are large differences even when we create dozens of rep lines.

Why should this be so? As a motivating example, consider valuing and estimating the risk of a portfolio of 200 equity put options (rather than mortgages). Each option has a different strike price, a different expiration date and is tied to a different underlying firm's equity. Even if an analyst knows the *average* strike price, the *average* expiration date, the *average* historical volatility, etc., it would be quite difficult for her to estimate the value of this portfolio of options using only this information. The analyst would be similarly challenged in attempting to run a stress test (e.g., the S&P 500 rises by 10%) based on such summary information. Examining an historical time series of the option portfolio's value over a five year period would provide only limited insight into how it might behave in the future.

Analyzing a portfolio of mortgages based on portfolio summaries is a similar exercise, given the various options embedded in residential mortgages, such as those relating to mortgage insurance or rate resets as well as the more fundamental prepayment (call) and default (put) options. Because of the non-linear behavior of each asset's prepayment, default and severity processes, the interaction of these processes with each other, and the heterogeneity of typical mortgage portfolios, evaluating a portfolio based on summary statistics alone can be challenging and it is not surprising that analyses done using aggregate methods can produce less detailed results than those performed using loan-level approaches.

This implies that analyses done using aggregate data require the analyst to make a greater number of assumptions under different scenarios and to perform a greater number of "what-if" analyses in assessing the credit risk of mortgage portfolios.

However, the historical reliance by some on aggregate rather than loan-level analyses for mortgage portfolios may reflect market conventions and a perception that the differences implied by the two analytic approaches are only modest. The events of the recent years have provided new evidence that this may not be the case. The use of aggregates may also reflect the historical difficulty in obtaining loan-level information in many markets, where for example, disclosure of such data may not be a market convention or regulatory requirement. (This is still the case in many domiciles and for many asset classes.) It may also be due to the fact that the study of the impact of aggregation effects is itself an evolving area of research, and a number of key results have only been established in the past several years. Finally, this study is, to our knowledge, among the first to address the question of the effects on mortgage analysis of loan-level vs. aggregate treatments over practically relevant horizons in detail[5].

While our findings have direct implication for both whole-loan risk management and for the evaluation of the credit risk of tranches of RMBS securities, we stress the preliminary

[5] We note a recent study (Hanson, Pesaran and Schuermann, 2008) that focuses on the implications of heterogeneity on portfolios of *corporate* assets using a different research approach.

nature of our findings and note that a more refined analysis might extend the experimental design along a number of dimensions.

The remainder of this paper proceeds as follows: In Section 2, we present some empirical results using historical mortgage data. In Section 3 we outline our experimental design and describe the data aggregation approaches and models we use. In Section 4, we present empirical comparisons of the two methods using a detailed model of mortgage portfolio credit risk. We also present examples of how this behavior impacts the analysis of RMBS tranches. Section 5 provides some analytic background on aggregation bias and discusses some useful results from the literature. We discuss our results in Section 6.

2 Motivating empirical examples

We argue that loan-level analysis provides a more detailed assessment of the credit risk of a mortgage pool, particularly for non-homogenous portfolios. While there are theoretical conditions under which performing analysis at the aggregate level may be preferred to analyzing at the individual level, it can be difficult to find instances in which those conditions are met *in practice*, and this is our experience.

To motivate our work, consider an illustrative example where heterogeneity in a single risk factor results in substantially increased credit risk. We construct the example using a sample of Prime jumbo loans drawn from a broad cross-section of loans underlying RMBS securitizations that were originated in 2006. The average three year default rate for the entire sample is 3.07 percent. The average FICO score is 742 and the average CLTV is 76. (By examining the performance of these loans over the period of the recent crisis, we are able to incorporate newly available data on the effects of loan- vs. pool-level representations during periods of economic stress.)

Next, we select a subset of loans with FICO scores in the range [750, 775) and bucket the loans by CLTV bands. We report the corresponding historical three year cumulative default rate, in percentage points (Table 1).

Table 1. Distribution of default rate for different CLTV buckets for loans with FICO score in the range [750, 775) demonstrates substantial heterogeneity

CLTV			
Low <70	Medium [70,80)	High [80,85)	Very High >=85
0.51	1.49	1.68	4.00

To illustrate how heterogeneity matters, we now imagine constructing two portfolios with almost identical *mean* CLTV but with varying degree of *heterogeneity across* CLTV and use this stylized case to show how the credit risk for these two portfolios would be markedly different – even though these two portfolios would present almost identical summary credit risk measures if analyzed using aggregate pool-level methods.

We construct the stylized portfolios by drawing half the loans in each portfolio from one bucket and half from another.

- In the first portfolio, which we will call the *CLTV-homogeneous* portfolio, half the loans were randomly chosen from the Medium CLTV bucket and half the loans from the High CLTV bucket.
- In the second portfolio, referred to as the *CLTV-barbelled* portfolio, half the loans were randomly chosen from the Low CLTV bucket and the other half from Very High CLTV bucket.

The three year default rate for the CLTV homogeneous portfolio would be 1.59% (= [0.5 × 1.49] + [0.5 × 1.68]). The corresponding default rate for the CLTV-barbelled portfolio would be 2.26% (= [0.5 × 0.51] + [0.5 × 4.00]) – more than *40% higher* than the CLTV-homogeneous portfolio. For completeness, we note that the average FICO score would be the same for the two portfolios since all four buckets have the same average FICO score and the average CLTV for the CLTV-homogeneous portfolio would be 77.5, versus 75 for the CLTV-barbelled portfolio. So the barbelled portfolio would produce a significantly higher default rate despite having no difference in mean FICO and a *lower* average CLTV.

Clearly, the default rates did not "average out" due to the nonlinear relationship between CLTV and default. This simple example demonstrates that the difference in credit risk estimation for mortgage portfolios when using an aggregate approach can be significant (in this example, around 40%), suggesting a practical need to model credit risk on a disaggregated loan-level basis. This behavior results, in part, from Jensen's Inequality[6].

This simple example may have been *too* simple in that it used only a single factor. A reasonable alternative approach might be to include *multiple* aggregate factors in the analysis. For instance, rather than using just the mean CLTV in isolation, we might instead consider using *both* the mean CLTV *and* the mean FICO. We explore this example next.

To do so, we extend the previous example by breaking out all loans by both FICO and CLTV and then calculating the corresponding three year default rates, as shown in Table 2.

[6] Recall that Jensen's Inequality is a mathematical result which implies that for non-linear functions, the mean of the function (e.g., the prepayment rate) will have a different value than the function of the mean. In, say, prepayment terms, this implies that the average prepayment rate for a pool of loans will be different than the prepayment rate for the average of all loans. (This can be verified trivially: if the function were $f(x) = x^2$, then for a pool of two loans with values of x of 2 and 8, the mean of the function of x is $(2^2+8^2)/2 = 34$, while the function of the mean of x is $[(2+8)/2]^2=25$)

Table 2. Joint distribution of FICO and CLTV demonstrates substantial heterogeneity in default rates

		CLTV			
		Low <70	Medium [70,80)	High [80,85)	Very High >=85
FICO Score	Low < 710	2.39	4.95	5.53	9.71
	Medium [710,750)	1.01	3.22	3.50	7.01
	High [750,775)	0.51	1.49	1.68	4.00
	Very High >= 775	0.13	0.72	0.91	1.84

As seen from the table, both FICO and CLTV are important factors in determining credit quality. The default rates vary dramatically across loans with different FICO scores and CLTV levels. As expected, default is higher for higher CLTV loans and it is also higher for loans to borrowers with lower FICO scores.

It is clear from the table that knowing that, say, the *average* CLTV for a portfolio is less than 70% (first CLTV column of data in the table) does not say very much about the default rate of the portfolio, which might range from about 0.13% to around 2.39% depending on other factors. Similarly, knowing that the average FICO score was in the range [750,775) (third FICO row of data) tells us only that default rates might range from 0.51% to 4%.

We next again imagine constructing two stylized portfolios:

- The first portfolio consists of an equal number of loans from the two dark shaded buckets (Medium CLTV, High FICO and Medium FICO, High CLTV).
- The second portfolio consists of an equal number of loans from the two lightly shaded buckets (Low CLTV, Very High FICO and Low FICO, Very High CLTV).

The mean FICO score, mean CLTV, and the three year realized default rate for each portfolio are summarized in Table 3, below. It is clear from the table that although the CLTV of the second portfolio would be *lower* than that of the first portfolio and the FICO for the second portfolio would be only very slightly better than that of the first portfolio, the default rate of the second portfolio would be almost *twice* the default rate of the first portfolio.

Table 3. Three year default rates for a portfolio of loans from the dark shaded buckets and a portfolio of loans from the lightly shaded buckets

	Mean FICO	Mean CLTV	3 yr default rate
Portfolio A – homogeneous	746	77.5	2.55
Portfolio B – barbell	738	75.0	4.92

This stylized example assumed that each portfolio was drawn identically from of its two respective buckets. Real portfolios will exhibit some degree of variance due to the variance within each bucket. To evaluate the degree to which our stylized results accord with more realistic portfolios, this we next constructed two portfolios, each of which has approximately the same average FICO and CLTV, but which have different degrees of heterogeneity in both risk characteristics.

- The first portfolio, which we call the *uniform sample portfolio*, consists of 2000 loans sampled uniformly from the *entire data set*.
- The second portfolio, which we construct as a barbell portfolio, has one-half of its loans drawn from the Low FICO, Very High CLTV bucket and half the loans drawn from the Very High FICO, Low CLTV bucket. The lightly shaded cells represent the loans in this barbell portfolio.

These portfolios are summarized in Table 4 below.

Table 4. Three year default rates for a randomly selected portfolio (A) and a portfolio with different joint distribution of underlying factors (B)

	Mean FICO	Mean CLTV	3 yr default rate
Portfolio A – uniform sample	742	76	3.07
Portfolio B – barbell sample	738	75	4.92

Note that though the average FICO and CLTV are almost identical, the three year default rate for the barbelled portfolio is 4.92%, which is *60% higher* than the uniform sample portfolio. For these two portfolios, the implication is clear: in this case, aggregation masked some important credit information.

To get a sense for how general the results of Table 4 were, we randomly created 10,000 portfolios of 2000 mortgages each, using the two approaches.

Although the average FICO and CLTV for the portfolios was practically the same, in over 99 % of the portfolios constructed in the "barbell" fashion the default rate was *higher than the maximum* default rate of the uniform sample portfolios. This suggests that credit risk of a portfolio depends on the joint distribution of the risk factors, as rather than the simple means of the risk factors.

This behavior is also due to a form of non-linearity involving aggregation bias, which we discuss in more detail in Section 5. As this example illustrates, examining multiple

aggregate factors may still lead to a masking of the detailed interaction effects between the different risk factors, which may explain a substantial portion of the risk[7].

Our central observation is that the *joint* distribution of risk factors determines the credit risk of the pool and that the heterogeneity in this joint distribution provides important information. Even if we know the marginal distribution of important risk factors such as FICO and CLTV, in general, we can get more information about the credit risk in the portfolio by examining the loan level data.

In principle, with sufficient data, it might be possible to capture the appropriate levels of granularity within a mortgage pool with pool-level summaries. One possible approach is to exhaustively describe all conditional relationships within the data. For example, if there were k factors in the model each with m levels, we would create $m \times k$ cohorts. However, adequately including all such interactions would require so much data as to be impractical in most cases. Consider how Table 2 might look if it were extended to include dimensions for the quality of loan documentation, property type, and loan type along with the zip code of the property.)[8]

3 Experimental design

In Section 2 we provided some empirical evidence suggesting that inferring the default rate based on summaries of the loan characteristics of a portfolio alone provides only partial information. This is largely due to the heterogeneity of the typical mortgage portfolio, the non-linear relationships that relate default probability to loan characteristics and the importance of interactions among the various characteristics.

In order to provide some sense of the differential impact on *forecasts* of different levels of aggregation, we conduct a series of experiments to forecast losses on mortgage portfolios. We create cohorts, calculate the average characteristics of the loans in each cohort and run these averages through models of credit portfolio behavior. We repeat this using the detailed loan-level information. We then examine the differences in the model forecasts when using the aggregate versus the loan-level methods.

We describe how we construct these cohorts (rep lines) from the raw loan-level data and then go on to describe our experiments.

[7] This also does not appear to be an artifact of the use of securitized mortgage pools for the experiments. We obtained substantially similar results, not reported here, when we instead selected the full mortgages portfolio (securitized and unsecuritized loans) of a large mortgage originator.

[8] In general, the number of loans required to populate all cells in such a table with reasonable confidence bounds would likely run into the millions. See: Stein (2006) for a discussion.

3.1 Constructing rep lines

We construct rep lines using an approach similar to that used in industry. We also create rep lines at different levels of detail in order to examine the sensitivity of our results to increased granularity in the rep lines. (This would only be feasible to a user with access to the loan-level data itself.)

For each pool in our sample, we construct rep lines in two steps:

1. First, we partition the loans in the pool into one or more cohorts. For example, if we are aggregating by geography, we create a cohort of loans that for each of the 50 states (plus Washington, DC and some territories) in which the loan properties in the portfolio are located.

2. Next, for each cohort constructed in step 1, we calculate the central tendency of all loans in that cohort for each loan attribute. (For example, to calculate the CLTV for a cohort, we estimate the mean CLTV for all loans in the cohort.) Loan attributes may be categorical variables such as mortgage type, property type, and state; or continuous variables such as CLTV and interest rate. For categorical variables, we used the cohort mode as a measure of the central tendency. For continuous variables, we used the arithmetic mean. We then use these aggregates to construct a *single representative loan* whose attributes are the central tendencies of the attributes of the loans in that cohort. This single representative loan for the cohort is called a rep line.

We use different dimensions to partition the pool to obtain different levels of granularity. Various rep line approaches differ in the level of detail used to define the cohorts. This also impacts the number of rep lines used in the analysis.

For our experiments, we formed rep lines in two ways:

Grand Average Rep lines: In this method, we use a single cohort comprised of the entire pool. This results in one rep line per pool.

Geographic Rep lines: To capture the effect of geographic features as well as some of the conditional behavior (e.g., concentrations of high FICO in certain regions) we created one cohort for each geographic state in the US. Thus, for each pool, each rep line contains state-specific averages of the loan attributes. This resulted in an average of about 32 rep lines per pool.

3.2 Mapping pools to rep lines

Once the rep line has been constructed for each cohort in the pool, we use the following method to calculate the expected loss for the pool of loans.

For each rep line, we determine a prepayment, default, and severity vector. These vectors are generated by running the loan through our model using a single macro-economic scenario, namely, the average economic forecast.

Each loan in the portfolio is then mapped to the prepayment, default and severity vectors for the corresponding rep line[9]. Because there is no stochastic behavior in this approach, each loan defaults (prepays) fractionally in each period. The fraction of the outstanding balance that is defaulted (prepaid) is proportional to the default (prepayment) probability of the corresponding rep line. For example, if we had a Texas loan represented by a Texas rep line, and the default rate in period 5 were 1% for that rep line, 1% of the par of the loan would be defaulted and the severity would be taken as the severity for that rep line in period 5. Note that in this setting, a specific economic scenario must be chosen to generate the vectors for each rep line. This approach is analogous to the way in which many market participants analyze RMBS transactions using predefined prepayment and default timing and level assumptions (except that we perform this at more detailed levels of stratification in some cases). (We discuss some implications of scenario analysis versus analysis of the full loss distribution in more detail in Appendix A.)

3.3 The model

We use a model of loan portfolio credit risk to assess the performance of the different aggregation approaches versus the full loan-by-loan representation. The model and its performance are described in detail in Stein, Das, Ding and Chinchalkar (2010).

The model analyzes mortgage portfolios in four steps. First, it generates trajectories of economic scenarios at a quarterly frequency over a ten year horizon. Next, for each loan in the portfolio, econometric models calculate quarterly default and prepayment probabilities for each quarter over a ten-year period as a function of loan-specific and economy-wide factors. Given these probabilities, the software then simulates default events, prepayment events, and loss given default and then for each trajectory, the model aggregates the simulated losses across all loans in the portfolio. Next, these simulated portfolio-level losses are aggregated across all trajectories to estimate the distribution of pool-level losses. Historical economic data for the simulations are updated quarterly.

The econometric models of prepayment, default and severity are estimated at the loan level and are related through common dependence on macroeconomic as well as loan specific factors. The macro-economic factors used in the loan-level simulation are generated at the national-level, state-level and MSA-level using econometric models developed for these factors. When simulating economic scenarios (rather than using fixed stress scenarios), the simulation is conducted over a 10-year horizon at a quarterly frequency using 10,000 equally-weighted simulated economic paths. This produces default rates, prepayment rates and severity for each loan in the portfolio for each economic scenario.

Interested readers can find a more detailed discussion of the models used in Stein, Das, Ding and Chinchalkar (2010).[10]

[9] This is analogous to par-weighting each rep line based on the total loan amount in each cohort.

[10] It could be argued that were a model specifically estimated at the aggregate level, it could perform better. In general it is difficult to say, though we discuss the literature on challenges to using this approach in Section 5. We also discuss a simple experiment we conducted to evaluate the degree to which fitting at the aggregate level might change our conclusions. While not

3.4 The experimental design

We performed two types of experiments. In one set, we use the prepayment, default, and severity models to calculate the expected loss for each portfolio under the base-case economic scenario from Moody's Economy.com, while in the other we perform a stress test on each portfolio. We use the "Complete Collapse" economic scenario produced by Moody's Economy.com as an example of the stress test. A brief description of this stress test scenario is provided in Appendix C.

These experiments were performed on 100 RMBS pools of loans from various trustees.

4 Empirical results

4.1 Average differences by method

Table 5, below, shows the results of two of the experiments on the set of 100 pools of Prime RMBS transactions. The table shows the average of the model's estimated expected loss (EL) across all pools for the average economy and the MEDC Complete Collapse scenario (MEDC Scenario 4).

Table 5 Summary statistics for results of analysis different methods on 100 RMBS pools

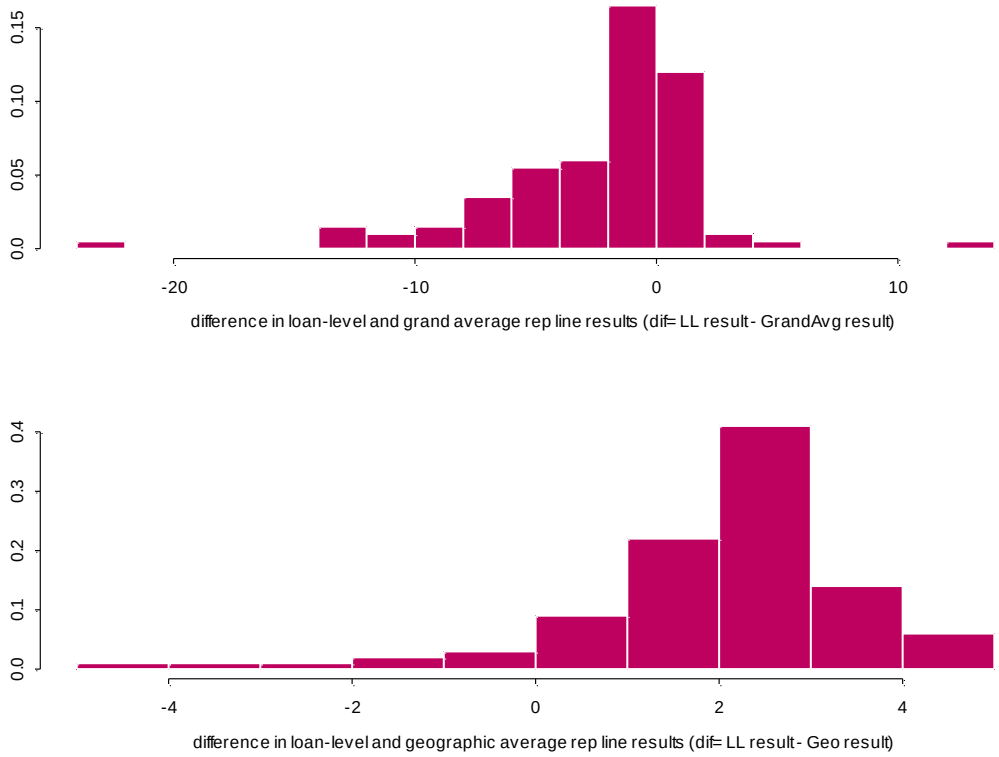
	EL-Average Economy	EL – MEDC Scenario 4
Full detail loan-level	13.0	20.7
Grand Average	15.3	24.7
Geographic Averages	11.0	19.4

We can observe that on an average the aggregation methods do not produce results that agree closely with the results of the full analysis (or with each other)[11]. This is not surprising, given the results of the historical tests. What may be surprising, however, is that in addition to producing results that are different than the loan-level analysis, the rep lines do not produce average results that are consistently conservative or liberal in comparison with the full loan-level analysis. This can be seen from the histograms in Figure 1 which show the differences between the rep line results and the loan-level results for each pool.

comprehensive, our experiment suggests that the conclusions we present would not change materially. A brief description of the experiment can be found in Appendix B.

[11] In all cases the means of the distributions were statistically different at high levels of significance, i.e. $p<<0.0001$ in a two tailed paired t-test.

Figure 1 Histograms of bias in various methods



There is no reason, in principle, why the rep lines should produce outcomes that are always higher or lower than the full analysis. For example, in pools in which the average loan produces a high expected loss, the outliers are the good loans which get excluded (or under-represented) in the averaging. On the other hand, for pools where the average loan produces a low expected loss, the outliers are the bad loans which get excluded (or under-represented) in the averaging.

We also observe that the losses under the tail scenario (MEDC Scenario 4) can vary substantially. This may be another result of Jensen's Inequality, though it is not directly related to our discussion here. We discuss this in more detail in Appendix B.

4.2 Examples of where portfolio-specific differences are far greater

The average differences in Table 5 are somewhat misleading. They imply a much closer agreement between the various methods than may actually be the case in some instances. To give some sense of how different the results can be using these different approaches, Table 6 shows the results of the different methods on a few pools. These differences are more pronounced. As before, the differences are not consistently higher or lower than the loan-level results, even within the same method.

Table 6 Expected loss for several pools using different aggregation methods

	Full detail loan-level analysis	**Grand Average**	**Geographic Average**
Pool 1	**1.26**	0.31	0.66
Pool 2	**0.25**	<1bp	0.05
Pool 3	**2.53**	0.63	0.39
Pool 4	**4.00**	4.35	1.86
Pool 5	**7.56**	8.72	5.52

4.3 Validation using realized data

We now consider the accuracy of the three methods in forecasting the default rate for a set of loans *ex post*. For this experiment, we consider a set of prime jumbo loans outstanding at different points in time and observe the default rate for that set of loans over the next three years. We then run the model using loan-level data as well as the two methods of aggregation *using the realized economic path* and determine the predicted default rate over the next three years. A comparison of the predicted default rate with the actual default rate helps us evaluate the accuracy of the model conditional on the economy. This approach separates the prediction of the economic path from the ability to relate the path to an observed outcome. (Stein, Das, Ding and Chinchalkar, 2010).

Table 7. A comparison of the realized three-year default rate and the three-year default rate predicted by the three methods

	Realized	Loan-level	Grand Average	Geographic Average
2006-Q1	**5.1**	5.2	1.8	2.9
2006-Q2	**5.6**	5.6	2.3	2.0
2006-Q3	**6.7**	6.9	2.8	1.9
2006-Q4	**7.3**	7.7	4.9	3.4

From the table, we can see that the default rate predicted using aggregate data is markedly lower than the realized default rate under both aggregation methods, whereas the loan-level analysis predicts similar values.[12] (Note that these results are based on fairly large samples. As the size of pools decreases, the variability of the results naturally increases.)

[12] A more detailed summary of the validation of the loan level models can be found in (Stein, Das, Ding and Chinchalkar, 2010).

4.4 Tail risk: comparing losses on structured RMBS tranches

We finally consider an example of tranches of a prime RMBS deal analyzed using these methods of aggregation. In this example, we generated cashflows for the underlying mortgage portfolio under each of the aggregation methods using both the mean economy and the MEDC "Total Collapse" scenario. We then used these cashflows to estimate losses on tranches of a structured transaction in the Structured Finance Workstation™ produced by Moody's Analytics.

For this example, we created a pool of approximately by selecting 1000 loans originated in 2009. The transaction was a cashflow structure consisting of five tranches. The notional amount for the senior-most tranche (A-1) was 96% of the outstanding pool balance, whereas the notional for each of the lower four tranches (A-2 through A-5) was 1% of the pool balance.

The table below shows the expected loss, in percent, using the different methods of aggregation for the average economy and the MEDC "Total Collapse" scenario.

Table 8. Expected loss (%) for the bottom tranches of an RMBS transaction

		Loan level	Grand Average	Geographic Average
MEDC-0 "Base case"	A-4 Tranche	-	-	-
	A-5 Tranche	7.7	0.6	2.4
MEDC-4 "Total Collapse"	A-4 Tranche	52.1	-	-
	A-5 Tranche	99.9	16.2	50.1

Table 8 shows a substantial difference between the expected loss generated by the methods of aggregation and the expected loss from the loan-level analysis.

We note under some scenarios, the expected loss for a tranche is zero, even in the loan level case. This reflects the use of a single scenario in the analysis. Were we to perform full simulation[13] the expected losses would be nonzero for all tranches. See Appendix A for a discussion.

5 Why aggregation may be misleading: an informal survey of the literature and discussion of implications for modeling mortgages

The crux of the motivation for loan-level analysis stems from the observed non-linear properties of mortgages, the heterogeneous nature of certain mortgage portfolios and some mathematical results and statistical theory, all of which generally fall into various

[13] When performing loan-level analysis, we run a full Monte-Carlo simulation and produce cashflows at the pool level for 10,000 economic scenarios. For each economic scenario, we run the cashflows through the waterfall and determine losses on each tranche. For each tranche, the average of these losses produces the expected loss.

broad interpretations of Jensen's Inequality, which manifest themselves in a number of ways in mortgage analysis. We provided examples of these effects in Table 1-Table 4 using historical data that suggested that the processes that drive mortgage default exhibit these effects with the result that inferences based on aggregations of historical data may imply quite different risk levels than inferences based on more detailed analysis of the loan data. We also provided examples in Table 5 through Table 7 that showed that models of these processes similarly demonstrate the impact of the non-linearity and heterogeneity. One implication is that that forecasts of portfolio losses based on aggregate data would typically be quite different than forecasts based on loan-level data.

As it turns out, this topic has been debated extensively in the statistical, economics and political science literature to such an extent that terms of art have evolved as shorthand for its various forms. For example, in the social sciences, a variant of this phenomenon may be termed *ecological fallacy* while in economics other versions of it may be termed *aggregation bias*. This debate continues. However, only recently have some of the newer results provided normative implications.

Though Jensen's Inequality is likely familiar to most readers from introductory mathematics classes (see Footnote 6) the literature on the statistical implications may not be. In this section, we describe briefly some of this literature and provide a sampling of some of the topics and results that have come out of the study of these phenomena. A fuller literature review can be found in a number of articles we cite and their references.

5.1 Micro-relationships from aggregate relationships

The most basic form of aggregation problem has been termed the *ecological fallacy* and has been well studied in the political science and epidemiology literature. The kernel of the concept is *that group summary statistics* (i.e., means within a specific subset of the population) typically cannot serve as proxies for the attributes of *individuals* within the group. Said differently, the relationship between two aggregate measures is not generally the same as the relationship between the equivalent two micro-measures. This may also be extended to instances in which aggregate behaviors cannot be directly observed (e.g., voting patterns or crime rates) and are predicted using aggregate data (Gelman, *et al.*, 2001).

One of the first mathematical explanations for this phenomenon was discussed by Robinson (1950) who concluded

> The relation between ecological and individual correlations which is discussed in this paper provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot. While it is theoretically possible for the two to be equal, the conditions under which this can happen are far removed from those ordinarily encountered in data. From a practical standpoint, therefore, the only reasonable assumption is that an ecological correlation is almost certainly not equal to its corresponding individual correlation. (Robinson, 1950).

This effect was subsequently studied in more detail by a number of authors and special cases in which the effect could be minimized were proposed. Theil (1966) provided a hierarchy of objectives in evaluating aggregation. The author also demonstrates three examples of applying the hierarchy to real data. For purposes of our discussion, the most

relevant of the three objectives is the author's first: given a set of reasonable criteria for using aggregation rather than individual analysis, define the set of all methods that meet these criteria (and select among them). The author concludes that for all real-world cases examined, the set of methods that would satisfy these criteria is empty, i.e., for the practical problems considered no method provides a reasonable substitute to individual level analysis.

5.2 Modeling the aggregate vs. aggregating micro-models: the linear case

A more relevant stream of the aggregation literature relates to the desirability of using individual level vs. aggregate data to estimate the *aggregate* outcome (example, using aggregate pool statistics to estimate the default rates for a pool of loans) when one or both of the aggregate and individual models and data are subject to error. Here much of the discussion revolves around the degree to which estimation error on *individual* observations and models, when aggregated, is higher than the error resulting from using *aggregate* data. In the economics literature, this is sometimes termed *aggregation bias*.

The majority of the early literature focuses on (often univariate) linear models, and thus may be of only passing interest in the mortgage setting, where nonlinearity is prominent. This focus on linear models reflects, in part, their tractability relative to non-linear approaches. It may also reflect the historical context (much of this work was done in the 1960s and 1970s, when data processing and analytic software was more limited than it is today and large-scale estimation of non-linear models was more cumbersome to perform). More recently, the literature has been extended to accommodate non-linear and multivariate models.

In one of the earliest analyses, Grunfeld and Griliches (1960) demonstrate that in linear settings in which there is substantial error in estimating the micro-models, aggregation may produce a benefit which may be sufficient to offset or exceed the aggregation error. The key assumption for this result to hold is that a model builder is either unable to get sufficient data to model well at the micro-level or the modeler is unable to accurately capture the underlying individual behavior at the micro level to produce a micro-model without large errors.

Ainger and Goldfeld (1974) extend this work and provide an analysis for the stylized univariate linear case. The authors outline specific circumstances under which the linear single factor aggregate model should be theoretically preferred to the micro-level model. They conclude that for prediction, if the coefficients of the individual-level model are the same across the units being aggregated the individual model will be superior. The same conclusion holds in general when the coefficients are unequal, however *some exceptions exist*. Because the model the authors examine is abstract relative to models used in practical mortgage analysis, these technical results are not normatively useful in our context, but do provide a sense of how the problem may be formulated. Firebaugh (1978) provides a different analysis, and derives a rule for determining in the multivariate linear setting when aggregation would provide an acceptable representation of the individual level behaviors, but again concludes, that without individual level data, the researcher "cannot determine, empirically, whether the data conform to the rule." Subsequently,

Greenland and Morgenstern (1989) demonstrated that even this simple rule is not adequate for evaluating non-linear models.

5.3 Modeling the aggregate vs. aggregating micro-models: the non-linear and heterogeneous cases

Much of the discussion in the present paper addresses aggregation of non-linear processes. While most of the early literature on aggregation is limited to the linear case, in the early 1980s, perhaps as a result of the increasing sophistication of the non-linear models used in finance and the availability of computing resources to estimate them, researchers began to focus on aggregation of non-linear micro-models. Kelejian (1980) takes up non-linear aggregation of stochastic functions and concludes that, in most realistic settings, it is unlikely that the macro-model can recover the micro-model. Further, it may often be the case that the prediction of the optimal (in an estimation sense) macro-model is inferior to other macro-models that are less well specified. The author recommends evaluating macro-models based on predictive power rather than their specification.

More recently van Garderen, Lee and Pesaran (2000) explicitly study the topic of aggregation of non-linear micro functions for prediction. The authors' basic conclusion is that except in the special case in which the aggregate and micro-equations produce identical results, "…if reliable disaggregate information is available, it should be utilized in forecasting, even if the primary objective is to forecast the aggregate variables." Thus, much of the discussion revolves around the *reliability* of the macro- vs. micro-level data and the models estimated on them. The authors propose a model-selection test for choosing between macro- and micro-level models and demonstrate the approach on two common non-linear economics problems. In both cases the micro-model is preferred. Importantly, in order to perform the test in van Garderen, Lee and Pesaran (2000), the modeler requires the full micro-data set.

Hsiao and Fujiki (2005) extend this work by comparing the results of an analysis using both aggregate and disaggregate data to measure Japan's aggregate money demand function. The authors also propose a new test for selecting between aggregate and micro-level models. In the authors' empirical analysis, the aggregate model and the disaggregate-models *lead to opposite conclusions* regarding a number of economic effects. They examine the differences in these predictions and attribute them to the heterogeneity across micro-units concluding (perhaps too strongly) that "…the prediction of aggregate outcomes, using aggregate data is less accurate than the prediction based on micro equations and policy evaluation based on aggregate data ignoring heterogeneity in micro units can be grossly misleading."

Blundell and Stoker (2005) provide a review on the state of the art on the issue of aggregation when there is heterogeneity across micro-units. The authors also discuss three specific examples of aggregation in the non-linear context. They focus only on the cases in which the non-linear functions of the micro-relationships are explicitly known. The authors show that in the case of basic non-linear functions, for which the form is explicitly known, it is feasible to accommodate aggregation in a non-linear setting,

provided the functional form of the aggregate is modified sufficiently to accommodate the specific form of the micro-relationships, and *provided the modeler is able to accurately estimate the probability distribution of the heterogeneous attributes* (i.e., using micro-data). The non-linear forms of the micro-functions considered in Blundell and Stoker (2005) were simple ones and the relationships they sought to explain were in two or three factors. When the underlying functional form is known and the micro data is available, the authors recommend hybrid approaches to integrating micro and aggregate data. While optimistic, the authors conclude that general solutions to the aggregation problem are still elusive:

> Heterogeneity across individuals is extremely extensive and its impact is not obviously simplified or lessened by the existence of economic interaction via markets or other institutions. The conditions under which one can ignore a great deal of the evidence of individual heterogeneity are so severe as to make them patently unrealistic… There is no quick, easy, or obvious fix to dealing with aggregation problems in general.

5.4 Empirical results for aggregate vs. disaggregated analysis

In closing out our review of some of the literature, it is useful to observe that the vast bulk of the empirical research on aggregation that we have discussed focuses not on the aggregation of *micro-units*, such as mortgages, but rather on aggregation of *sub-components*, such as the local unemployment rate, which *are themselves aggregates*, though of smaller areas. The empirical results here are mixed.

For example, Miller (1998) compared the approach of aggregating county-level forecasts of total employment to produce the FEA (functional economic area) forecast to forecasts produced from the aggregate FEA employment levels directly, and found mixed results including a number in which aggregate forecasts out-performed the aggregation of disaggregate forecasts. In an application of aggregation of linear models, Hubrich (2005) finds that for linear time-series forecasts of euro area inflation, aggregating component time series forecasts did not necessarily improve on forecasts made at the aggregate level, though in some cases it did.

Similarly, Hendry and Hubrich (2006) provide a theoretical motivation for the benefits of combining aggregate and disaggregate information under the assumption that a modeler *uses complete information* through the current period (here complete information would include the loan-level data). However, their empirical results are mixed when they apply the theory to real data. Their analysis focused on forecasting euro area inflation and US inflation, again, using disaggregated inflation series.

For heterogeneous settings, the results of Hsiao and Fujiki (2005), discussed in more detail above, suggest that in dealing with disaggregated components, the aggregate results were inferior to those from the disaggregated approach.

5.5 Summary of literature and implications for modeling mortgages:

Though debate continues, some patterns appear to be emerging in the past several years. Summarizing, the literature seems to suggest that:

a) Aggregation in general should not be expected to *recover* micro-level relationships, particularly when the micro relationships are non-linear and/or heterogeneous;
b) Aggregation generally will not produce superior results to disaggregate analysis, unless the noise in the micro-level data is sufficiently high that it obscures the signal;
c) The expectation of superior performance for aggregate-models is lower for non-linear models than for linear models and for heterogeneous settings than for homogeneous ones;
d) Aggregation may produce superior results (empirical evidence is mixed) when the components *are themselves aggregates,* such as would be the case when component estimates of unemployment are considered in forecasting aggregate unemployment; and
e) Tests for selecting between aggregate or micro-level models typically rely on the tester having both the micro-level and aggregate-level data (in which case the superior approach may be selected as either can be derived from the micro-data).

This has a number of direct implications for modeling mortgage losses.

First, in considering individual mortgage behavior, the questions of non-linearity and heterogeneity become important ones. In cases in which the individual behaviors are non-linear and heterogeneous across individuals, aggregation holds less promise. In particular, it is unlikely that the aggregate model will accurately reflect the dynamics of the underlying process unless the underlying non-linearities are known functions and these functions are tractable analytically. In the mortgage setting, research suggests that the relationship between, e.g., default probability and loan factors is non-linear, and in some cases highly so, as are the relationships of the factors to each other (c.f., Stein, Das, Ding and Chinchalkar, 2010). This also suggests that loan-level models be specified to accommodate both single-factor non-linearity and their interactions.

Second, in the presence of strong non-linearity, while it is theoretically possible to achieve superior performance by using aggregate data under strict assumptions, the probability of doing so may be low, unless the data on the underlying mortgages is of particularly poor quality. Thus, substantial attention should be given to data cleaning and factor selection. For example, DTI is a factor which may be reported using a number of conventions and these vary substantially across time and mortgage originator. Such a factor, while theoretically useful, may not be appropriate for inclusion in certain models due to this observational noise. More generally, most mortgage data requires extensive error correction prior to use in modeling due to inconsistencies and data.

Third, to rule out excessive estimation and data noise, extensive validation of loan-level models is warranted not only across broad aggregate samples of data, but also for subsets of the data that may have different behavior than the general population. Stein, Das, Ding and Chinchalkar (2010) provide a number of examples of such validation.

Finally, we cannot generally expect aggregate models to provide guidance at the individual loan level. For example, we do not expect an aggregate model of pool behavior to provide guidance as to which loans are driving losses or which loans are most likely to prepay in the next year.

6 Discussion

Why should it be that aggregation at times reduces the information content in portfolio data to such a high degree?

Consider the case of estimating the mean loss (EL). In our tests we sometimes found the EL using the full detail to be higher than that produced by the rep line approaches. The economic intuition for this follows by considering that for prime mortgage pools, the EL is primarily driven by a (relatively) small number of bad loans. The typical mortgage portfolio is often constructed to provide diversification and thus exhibits heterogeneity in loan type, geography and so forth. This heterogeneity of the pool is not well represented when loan characteristics are aggregated. Said differently, for e.g., good quality portfolios, the average loan fails to capture the very bad characteristics of the few loans that drive the pool EL.

Aggregation can mask extreme observations. In case of categorical variables such as documentation type, the extreme observations (the very worst or very best loans) are often masked during the construction of rep lines, since the tails of the distribution do not contribute to the mode. For continuous variables such as LTV, although the outliers are represented in the average, because of the nonlinear relation between EL and LTV and the effect of Jensen's Inequality, as demonstrated in Section 2, the EL for the average loan may be lower than the average of the ELs for all the loans.

In both cases, this may be further exacerbated by the selection of a single scenario (macro-economic or default/prepay) to evaluate the mortgage pool. As we discuss in Appendix A, if most loans do not default in the "mean" economic scenario, but they do default in more extreme scenarios, the mean loss on the pool (its EL) will be greater than the loss on the pool under the mean scenario. (The converse is true for portfolios in which many loans default under the mean scenario.) Thus, when examining a specific stress case, the joint distribution of loan factors and the impact of non-linearities are more pronounced as the focus of these analyses is typically on the tail of the distribution, where the non-linearity is often most evident and this may not be fully captured by a single path.

We have provided empirical intuition for several reasons why methods of aggregation tend to produce results that do not agree with analysis done using the original loan-level data:

1. *Simple application of Jensen's Inequalities*: As we discussed in a number of places in this article, the portfolio loss levels are non-linear functions of the loan attributes. Jensen's Inequality implies that the EL for a pool of "average" loans will be different from the average of the ELs for the pool of loans. (Of course, the pool level EL will be the mean of the ELs of each loan).

2. *Aggregation bias*: In simple rep line construction, since aggregation is performed independently for each loan attribute, any information about the joint distribution

of the values of different attributes is lost. For example, suppose higher FICO loans have a low CLTV and lower FICO loans have a high CLTV in a given pool. When we average FICO and CLTV independently, we end up with one representative loan with an average FICO and an average CLTV. Information about the variation of FICO with CLTV is "lost". Therefore, we cannot distinguish this from the case where high FICO loans have high CLTV and low FICO loans have low CLTV. We showed this in examples given in Section 2.

This limitation exists even when information is available about the marginal distribution of each of the loan attributes. For example, suppose a full table of FICO frequencies for loans in a mortgage pool were available and a similar table of frequencies for the loan CLTV were also given. While this provides substantially more information than does the simple average, these frequency distributions do not provide information about the *joint* distribution of FICO and CLTV. So, in the example above, the pool containing high FICO loans with low CLTV and low FICO loans with high CLTV will have the same marginal distributions of FICO and CLTV as the pool containing high FICO loans with high CLTV and low FICO loans with low CLTV. The information would also not reveal, for example, whether the high CLTV loans were concentrated in, say, California, or were distributed more evenly across the country. Therefore, two pools could have identical summary statistics but very different loan-level characteristics. We also presented an example of this in Section 2.

3. *Masking of extreme values*: Pool performance may be driven by the very worst or very best loans which, by construction, are either not represented or under-represented in aggregates. This is especially true of categorical attributes such as property type and documentation type, where the mode of the distribution is insensitive to the outliers. This can be pronounced in cases in which an analyst is using one or a few paths to evaluate a pool (as is often done in industry) since masking of the defaults under adverse scenarios will also be present.

Before leaving this section we note that there are many cases in which loan-level data (particularly loan-level data containing borrower characteristics) is not available. This now happens less frequently in the analysis of US residential mortgages, but it is the norm for many other retail asset classes for which rep lines are used. In such cases the only viable approach to understanding credit behavior may be to estimate it at the pool-level and this does provide some useful guidance.

Even when loan-data is available, there may still be reasons a user would prefer to use rep-line analysis. For example, users wishing to perform cashflow waterfall analysis using tools for which there is no integrated loan-level simulator have little choice but to use rep lines. In some cases this approach, augmented with qualitative adjustments, may provide satisfactory analysis for such users.

It is evident from Table 6 that even if the pools are heterogeneous, there is merit to knowing the average values of key credit drivers like FICO and CLTV.

However, we can say a bit more about the information in aggregates.

Table 6 suggests that, in the absence of loan-level data, pool-level data may still provide useful clues to the credit quality of the different portfolios. The ranking of the pools based on aggregate data is similar in many cases to the ranking of the pools using the loan-level data. Thus, the pool-level data is able to separate, in broad terms, the better quality pools from the worse quality pools. It is not the case, though, that the aggregate data is able to rank pools in all cases, which is why there is a lack of agreement on the ordering of pools between the different methods.

It is notable that portfolio summary reports (including trustee reports that summarize collateral pools underlying RMBS) sometimes do not contain data that is refined enough to create the fine stratifications that we have created for our experiments. Rather, these summary reports often provide only a few averages and some descriptive geographical information. Thus, relative to the complexity of mortgage instruments and the heterogeneity of mortgage pools, the summary information upon which analysts must base their judgments in these cases is often sparse. For example, a typical summary report for an RMBS transaction might contain tables of portfolio averages or distributions for FICO, CLTV, loan rate, and so forth. In most cases, these summaries take the form of univariate (marginal) distributions or averages. Occasionally, a bivariate distribution may be given, e.g., average FICO in each of the pool's largest ten states.

The experimental evidence we have presented suggests that when loan-level data is available, it is informative. Of course, if loan-level data is available, an aggregate analysis may always be performed in addition to the more detailed loan-level analysis, if it is desired.

Finally, we note that many of these results are most evident during periods of economic stress. Prior to the recent years, it was considerably harder to observe the relationships between factors that demonstrated the strong heterogeneity that is central to the performance differentials between aggregate and disaggregate approaches.

7 Conclusion

We conducted a number of experiments with the objective of examining the impact of aggregation on the analysis of residential mortgage pools. We benchmarked a number of approaches to aggregating detailed loan-level information into smaller numbers of rep lines. Our research suggests that credit losses predicted based on analysis of loan-by-loan data may be materially different than the loss based on the aggregates of these factors.

Our results suggest that loan-level analysis may produce more refined results than pool-level or summary analysis. We find that this effect persists even when we further subset a pool into dozens of rep lines. The differences in analyses may be particularly pronounced for more heterogeneous portfolios.

These findings are consistent with statistical theory and we attribute them to three main effects: simple cases of Jensen's Inequality, aggregation bias, and the masking of extreme observations. These results impact mortgage pools because of the non-linear behavior of the individual assets' processes (prepayment, default and severity), the non-linear interaction of these processes with each other, and the heterogeneity of typical mortgage portfolios.

Loan-level analysis is not always feasible. In such situations aggregate approaches can provide a mechanism to understand the relative credit risks in mortgage portfolios. However, our findings imply that using the full detail for mortgage pool analysis, when it is available, may produce more realistic representations of loan behavior.

Current practice in the analysis of RMBS transactions at some institutions tends to favor the use of rep lines – at least during the evaluation of the cashflow waterfall of the bonds. This practice is also used for whole-loan portfolios by some market participants. A primary reason for this is that it can be exceedingly difficult to perform loan-level loss analysis within most waterfall modeling software tools, except on a scenario basis. Because of this limitation, analysts may rely on one or a few rep lines and a handful of stress scenarios.

The experimental results here suggest that where feasible, a detailed analysis of the collateral pool, at the loan-level, provides a richer representation of portfolio dynamics. This is true for whole-loan portfolios and for collateral pools underlying RMBS transactions, where the impact can be pronounced.

New tools that permit loan-level asset analysis in a simulation framework within a cashflow modeling software hold promise for reducing levels of aggregation bias in RMBS analysis.

8 References

1. Ainger, D. J. and S. M. Goldfeld (1974), Estimation and prediction from aggregate data when aggregates are measured more accurately than their components, *Econometrica*, **42**, 1, 113-134.

2. Blundell, R. and T. M. Stoker (2005), *Journal of Economic Literature*, **XLIII**, 347-391.

3. Firebaugh, G. (1978), A Rule for Inferring Individual-Level Relationships from Aggregate Data, *American Sociological Review*, **43**, 4, pp. 557-572

4. Greenland, S. and H. Morgenstern (1989), Ecological bias, confounding, and effect modification, *International Journal of Epidemiology*, **18**, 1, 269-274.

5. Gelman, A., D. K. Park, S. Ansolabahere, P. N. Price, L. C. Minnite (2001), Assumptions and model checking in ecological regressions, *Journal of the Royal Statistical Society, A*, **164**, 1, 101-118.

6. Grunfeld, Y. and Z. Griliches (1960), Is Aggregation Necessarily Bad?, *The Review of Economics and Statistics*, **42**,1, 1-13.

7. Hanson, S. G., M. H. Pesaran, T. Schuermann, (2008) Firm heterogeneity and credit risk diversification , *Journal of Empirical Finance*, **15**, 4, 583-612.

8. Hendry, D. F. and K. Hubrich (2006), Forecasting Economic Aggregates by Disaggregates, Working Paper No. 589, European Central Bank, February.

9. Hsiao, C., Y. Shen and H. Fujiki (2005), Aggregate vs. disaggregate data analysis—a paradox in the estimation of a money demand function of Japan under the low interest rate policy, *Journal of Applied Econometrics*, **20**, 5, 579–601.

10. Hubrich, K. (2005), Forecasting euro area inflation: Does aggregating forecasts by HIPC component improve forecast accuracy?, *International Journal of Forecasting*, **21**, 1, 119-136.

11. Kelejian, H.H. (1980), Aggregation and disaggregation of non-linear equations, In: Kmenta, J.,Ramsay, J.B. (Eds.), *Evaluation of econometric models*. Academic Press, New York.

12. Kramer, G. H., The Ecological Fallacy Revisited: Aggregate- versus Individual-level Findings on Economics and Elections, and Sociotropic Voting (1950), *The American Political Science Review*, **77**, 1, 92-111.

13. Miller, Jon R. (1998) *Journal of Regional Analysis and Policy* , **28**, 1, 49-59

14. Robinson, W.S. (1950). "Ecological Correlations and the Behavior of Individuals". *American Sociological Review*, 15, 3, 351–357

15. Stein, R. M. (2006), Are the probabilities right? Dependent defaults and the number of observations required to test default rate accuracy, Journal of Investment Management, 4, 2, 61-71

16. Stein, R. M., Das A., Ding, Y., Chinchalkar, S. (2010), Moody's Mortgage Metrics Prime: A Quasi-Structural Model of Prime Mortgage Portfolio Losses, Technical Document, Moody's Research Labs, New York

17. Theil, H. (1966), Alternative approaches to the aggregation problem , *Studies in Logic and the Foundations of Mathematics, Logic, Methodology and Philosophy of Science*, Proceeding of the 1960 International Congress, 44, 507-527.

18. van Garderen, K. J., K. Lee, M. H. Pesaran (2000), Cross-sectional aggregation of non-linear models, *Journal of Econometrics*, **95**, 2, 285-331

19. Zandi, M. M. and Z. Pozsar, 2006, U.S. Macro Model System, *Regional Financial Review*, Moody's Economy.com, West Chester, PA, 2006

9 Appendix A: A comment on the use of stress scenario probabilities for tail estimation

Though it is not the main focus of this paper, a common use of rep line analysis is to perform stress testing and scenario analysis. One application of such scenario analysis is to size tail risk. This is often accomplished by running an analysis using a stressed economic scenario that has an associated probability of occurring. (Such scenarios and probabilities might be produced by a third party or an in-house economics unit.) To round our discussion of the use of rep lines, in this appendix we present some comments on estimating tail probabilities using the stress-case approach versus using a full loss distribution.

Stress testing has many applications, particularly with respect to model validation, calibration and assumption testing. We differentiate the use of stress testing as a *qualitative approach* to understanding and providing a reality check for a model or portfolio on the one hand, from the use of stress testing in a *probabilistic setting* as a quantitative measure, on the other.

In the qualitative case, stress tests provide a means to associate concrete views on states of the world with model outputs and to evaluate the outputs for reasonableness. In the probabilistic case, however, the assumption that the probability of an outcome is equivalent to the probability of a tail loss typically does not hold.

Stress-test users sometimes confuse the probability of an *economic scenario* exceeding the stress scenario (e.g, 10% of all *economic paths* will be worse than scenario *x*) with the probability of a *portfolio loss* exceeding the loss in the stress case (e.g., 10% of all *losses* will be worse than the loss under scenario *x*). These are typically not the same since for a specific portfolio, there may be many scenarios that generate losses that are higher (lower) than those experienced under the stress scenario, due to the specific features of the portfolio. (Mathematically, $p(\textit{stress scenario}) \neq p(L > L_{\textit{stress scenario}})$.)

Imagine three 10-year home price stress scenarios:

A. *Slowdown in growth, but growth remains positive*: Home prices rise ½% each year over 10 years.
B. *Prices drop*: Home prices decline by 5% over five years, ending in year 5 at pre-decrease levels minus 5%. After year 5, prices rise at 4.5% per year.
C. *Prices drop severely*: Home prices decline by 25% over the first three years and then rise to pre-decrease levels minus 5% over the subsequent two years. After year 5, prices rise at 4.5% per year.

Clearly, scenario C (*prices drop severely*) is *generically* worse than scenario B (*prices decline*) since the peak-to-trough home price decline is more severe in scenario C than in B, and the two are identical after year five. Scenario A (*slowdown in growth, but growth remains positive*) would be viewed by many as the least stressful.

Now consider three portfolios of mortgages:

1. A portfolio of 3/27 loans (loans that pay a fixed coupon payment for the first three years and then convert to a floating coupon, with a typically higher interest payment).

2. A portfolio of 5/25 loans (loans that pay a fixed coupon payment for the first five years and then convert to a floating coupon, with a typically higher interest payment);
3. A portfolio of 7/1 loans (loans that pay a fixed coupon payment for the first seven years and then convert to a floating coupon, with a typically higher interest payment) that resets each year.

For simplicity, we will examine one dimension of loan performance: that related to payment resets and the borrower's ability to refinance to avoid resetting to a higher monthly coupon payment. We also assume for simplicity that borrowers prefer to only repay at the time their loan rate resets at the end of the fixed-rate period[14].

Consider now how these scenarios affect Portfolios 1 and 2. For both portfolios, Scenario A is the least disruptive from a refinancing perspective. As the coupon reset approaches, loans in both portfolios have realized positive equity growth and, if refinancing makes sense from an interest rate environment standpoint, positive equity will allow them to refinance in order to avoid increased coupon payments.

The other scenarios are less clear.

- For Portfolio 1 (3/27 loans), scenario C (*prices drop severely*) is far more challenging than Scenario B (*prices drop*) and refinance risk will be higher under C than B. This is true because at the very time the interest payments on the mortgage is due to reset to a higher rate in year 3, the borrower has experienced high levels of loss in home equity. Prices have dropped 25% since origination and for many borrowers, their mortgages will be "underwater." Thus, even though they would like to refinance, they may not be able to, due to the negative equity. In contrast, under Scenario B (*prices drop*) these same borrowers will have experienced a much smaller decrease in equity, making refinancing still sensible in many cases. For these borrowers, the sharp increase in coupon payments will be avoided.

- For Portfolio 2 (5/25 loans), both Scenario B and Scenario C affect losses similarly. This is because by the time the loans reset in year 5, home prices are the same level under both scenarios and they then move identically in both cases.

Now consider how the three scenarios affect Portfolio 3 (7/1 loans). In this case either one of the "bad" scenarios (Scenario B or C) is slightly *preferable* to Scenario A (*slowdown in growth, but growth remains positive*). To see why consider that when the loans in the portfolio are due for rate resets in year 7, the home prices under Scenarios B and C will have experienced continued growth at 4.5% per year, which, starting from a 5% decline in year 5, puts the home prices at about 3.75% over the initial value. In contrast, under Scenario A, the 10 year growth has been a bit slower with seven years of ½% growth resulting prices levels of about 3.55%. Thus, under Scenario A, the loans in Portfolio 3 will have experienced a bit less home price appreciation than under B or C.

[14] This is empirically not exactly correct. After month 12, there is often a modest increase in prepayments as borrowers with lower FICO scores take advantage of their newly established track record of mortgage payments to renegotiate.

The relationships are summarized in Table 9.

Table 9 Summary of implied reset risk on different portfolios under different stress scenarios

	Portfolio 1 (3/27)	Portfolio 2 (5/25)	Portfolio 3 (7/23)
Highest reset risk in	C	B or C	A
Lowest reset risk in	A	A	B or C

From the table, it is clear that regardless of the probability associated with a specific economic outcome, the probability that reset risk will be high or low depends on the structure of the portfolio being analyzed.

In this expository example, we only focus on the ability to prepay in order to avoid increases in monthly interest payments. Clearly, losses on real loans and loan portfolios are governed by a host of other behaviors which interact with each other in a myriad of ways, all of which can affect the ordering of losses under different scenarios for a specific portfolio.

We present a few examples in Table 10 that demonstrate this empirically on our data. We have selected a subset of the 100 mortgage pools that we used in our earlier analysis and run each using both the "Total Collapse" stress case and the full simulation. To simplify the presentation, we assume a user is interested in evaluating the amount of capital required to ensure that losses will be no greater that L with probability $1-\alpha$. This is commonly referred to as the "$1-\alpha$ value at risk level" or the "$1-\alpha$ VaR."

Table 10 Examples of stress scenarios on different pools

	Loan-level 95% VaR	EL under S4	Pct losses worse than losses under S4
Pool A	11.4	16.0	0. 1
Pool B	9.0	13.1	0. 1
Pool C	5.6	7.9	0. 1
Pool D	27.0	27.2	4.5
Pool E	20.2	20.3	4.6
Pool F	27.0	26.6	5.9
Pool G	38.6	33.9	15.5
Pool H	46.6	40.7	18.1
Pool I	44.2	34.5	25.5

From the table, it is clear that under the same scenario, the probability of losses exceeding those experienced under the scenario can vary greatly. This highlights the great difficulty in ordering economic scenarios from best to worst in general, and in translating the probability of a scenario's occurrence into a probability that losses will be greater than those under the scenario. (Note also that since scenario-based approaches produce a single path, differences will be observed in the ELs produced by simulation versus the mean scenario because the mean of the loss distribution is an average over all scenarios, rather than just the loss for the average scenario.[15]).

This underscores the benefit of performing both types of analysis.

- Scenario analysis provides a concrete, intuitive description of states of the world that might occur and the losses associated with those states under the model. This is valuable for understanding a model and a portfolio and for gaining intuition on the drivers of credit risk for the portfolio.

- Simulation-based analysis is one way to produce a detailed description of the loss distribution for a portfolio and gives information about the range of possible losses in a probabilistic setting. This can be useful in assessing capital requirements and for other portfolio management activities.

[15] To see this, consider the stylized example of a single mortgage with an LTV of 80%. In this example, assume further that the loan only defaults when its LTV rises above 80%. Imagine that in the average economic scenario, home prices are flat, so LTV does not rise above 80%. Thus, under the average scenario, the loan does not default and has a default rate of 0%. However, consider what happens in bad states of the economy: home prices fall and the loan defaults. If this happens in 1% of all cases, then the mean default rate across all scenarios will be 1%. Thus in the case of the average path, the fractional default approach under the average scenario would suggest a default rate of zero while the loan replacement approach would suggest a default rate of 1%. This happens even when the mean of the simulated paths is exactly the same as the mean economic scenario used in the fractional default case.

10 Appendix B: Fitting a simple model to aggregate data rather than loan-level data

A possible concern in our analysis is that we used a loan level model to test pool-level aggregates.

While it is beyond the scope of this paper to address this question in detail, to give some sense of the degree to which our conclusions might change if we performed forecasting based on a model *fit* at the aggregate level, rather than at the loan level, we fit two sets of simple models, using the two loan factors we discussed in Section 2, FICO and CLTV.

10.1 Data and design

We began by creating 200 heterogeneous pools of loans originated in 2006.

We fit our models on half of this data and used the models to predict pool-level losses on the other half. By restricting the model to a single period, the test is out-of-sample but in-time. Note also that by restricting our analysis to only those loans originated in 2006, we are effectively controlling for the impact of different economic environments, underwriting conventions and other time-varying factors such as changes in home prices, since all loans are exposed to identical historical economic periods and we predict into the same period. As prediction goes, this is a fairly tightly bounded experimental design.

We fit two types of models.

- Aggregate models were estimated by regressing the pool-level three year default rate on the mean FICO and mean CLTV for each of the pools. We estimate the model over 100 in-sample pools.
- The loan-level models were fit by pooling all of the in-sample data and regressing a three year default flag for each loan (1 = default within 3 years, 0 = no default) on each loan's FICO and CLTV.

We explored a number of functional forms for each model including linear and logit forms using either the levels of FICO and CLTV or a set of bins for FICO and CLTV to capture nonlinearity in the factors. In the case of the aggregate models, we also allowed for interactions between FICO and CLTV in levels and bins.

10.2 Summary of the results

In this experiment, the non-linear forms performed better in both aggregate and micro settings, with the micro-models dominating the aggregate models in all cases. A comparison of the error rates of the models yielded relative RMS error rates that were 85%-500% higher for the aggregate models than the micro models, depending on which two models were compared. The correlation of the predicted default rates with the actual default rates was also substantially higher for the micro models than for the aggregate models.

This experiment is highly stylized and is, by no means conclusive. It does, however, support the literature in that it suggests that in heterogeneous settings, it can be challenging to design aggregate models that perform better than micro-level models, even when the micro models are very simple ones and the domain is restricted substantially.

11 Appendix C: A summary of the MEDC Complete Collapse scenario

The stress scenario we use is provided by Moody's Economy.com. Additional details of the MEDC methodologies may be found in Zandi and Posner (2006). The following section is taken from Moody's Economy.com web site and describes in more detail the conditions assumed in this stress scenario.

> *With this depression scenario, there is a 96% probability that the economy will perform better, broadly speaking, and a 4% probability that it will perform worse.*
>
> The downside 4% scenario, "Complete Collapse, Depression Scenario," is caused by several factors. First, long-running restricted credit from banks prevents the consumer spending rebound from being sustained. Second, the debt crisis in Europe results in a deep recession there, causing U.S. exports to fall sharply. Third, the U.S. federal government reaches the limit of its resources to boost the economy, rendering it unable to prevent a deep economic slump. This scenario assumes that the effects of the 2009 federal stimulus proved to be only temporary. The recovery in the economy after mid-2009 essentially ends in the first half of 2010, as the census-related job creation in the first half of 2010 at best prevented a decline during that time. After June, the downturn accelerates and continues until the third quarter of 2011.
>
> In the housing market, foreclosure mitigation policies are unproductive. Businesses have little incentive to engage in investment spending because of the very low levels of capacity utilization, poor profitability, and the difficulty of obtaining capital. High unemployment and depressed income levels not only prevent consumers from obtaining access to credit but also cause them to return to a pattern of high precautionary saving and debt pay-down.
>
> Housing starts resume their decline and ultimately fall by 85% cumulatively from their 2005 peak. Although they finally bottom out in mid-2011, the increase is at a snail's pace for several years. House prices resume their decline, and the NAR median existing sales price ultimately falls cumulatively by 45% from its 2005 peak to the third quarter of 2012. Reduced household wealth, high unemployment, and the lack of credit cause consumers to pull back sharply on their spending. Unit auto sales average below 10 million throughout 2011 and 2012. Business investment falls throughout 2010 and 2011 and does not begin to recover until 2012.
>
> In the second recession, real GDP falls from the second quarter of 2010 until the third quarter of 2011, cumulatively declining by 2.7% peak to trough. On an annual average basis, real GDP growth is 1.3% in 2010 and -1.9% in 2011. The unemployment rate reaches a high of 15.1% in mid-2012 and remains in double digits until 2015. The extreme weakness results in consumer price deflation from mid-2010 through the end of 2011. (www.economy.com)