Data and Disaster: The Role of Data in the Financial Crisis

Louise Francis and Virginia R. Prevosto

Abstract:

Motivation. Since 2007 a global financial crisis has been unfolding. The crisis was initially caused by defaults on subprime loans, aided and abetted by pools of asset-backed securities and credit derivatives, but corporate defaults, such as that of Lehman Brothers, and outright fraud have also contributed to the crisis. Little research has been published investigating the role of data issues in various aspects of the financial crisis. In this paper we illustrate how data that was available to underwriters, credit agencies, the Securities and Exchange Commission (SEC), and fund managers could have been used to detect the problems that led to the financial crisis.

Method. In this paper we show that data quality played a significant role in the mispricing and business intelligence errors that caused the crisis. We utilize a number of relatively simple statistics to illustrate the due diligence that should have, but was not performed. We use the Madoff fraud and the mortgage meltdown as data quality case studies. We apply simple exploratory procedures to illustrate simple techniques that could have been used to detect problems. We also illustrate some modeling methods that could have been used to help underwrite mortgages and find indications of fraud.

Results. In both the Madoff fraud and the mortgage crisis a number of statistical tests could have been applied to uncover fraud and to provide a warning of the deterioration of the quality of mortgages underwritten.

Conclusions. Data quality issues made a significant contribution to the global financial crisis.

Keywords. Data, data quality, financial crisis

I. INTRODUCTION

In the eBook *Risk Management: The Current Financial Crisis, Lessons Learned and Future Implications*, published by the Joint Risk Management Section, a North American actuarial risk management organization, it was stated that, "The current financial crisis presents a case study of a financial tsunami…on what can go wrong. Its ramifications are far reaching and the lessons learned will be embedded in risk management practices for years to come."[1] In this publication a number of factors causing the financial crisis were identified: bubble behavior and mentality, liquidity problems, incentives (i.e., compensation), accounting disclosure issues, insufficient commitment to ERM, and flawed models. In addressing the issues raised with how models were used to price and rate securities based on mortgages, it was pointed out that there were data issues that had an impact on underlying assumptions embedded in the models. For example, the models assumed that housing prices would not decline over time. The following was quoted from a December 2005 report "the risk of national decline in home prices appears remote. The annual decline in HPA (i.e., housing price appreciation) has never been negative in the United States going back to 1992."[2]

[1] From the introduction of JRMS, 2008

[2] Schoolman, 2008

In addition to bankruptcies, forced buyouts, stock market, and bond value declines caused by the subprime meltdown, and the liquidity crunch caused by the crisis brought to light a number of large fraud schemes. One of the most prominent of these was the Ponzi scheme run by Bernard Madoff. What was most interesting was that a whistleblower, after examining data from one of Madoff's clients, first warned the Securities and Exchange Commission (SEC) about the fraud in 2000 and numerous times afterwards (Markopolos 2009).

Though it has not received a lot of publicity, poor data quality played a significant role in the global financial crisis that began to unfold in 2007. In this paper we will examine the role of data in the global financial crisis and in high-profile financial frauds. We will relate specific quality issues to the Actuarial Standard Board's ASOP No. 23 on Data Quality.

This paper will feature two data quality case studies:

1. An evaluation of data from a Madoff feeder fund[3] to illustrate a number of techniques that could have been used to determine that it was fake data.
2. A review of mortgage data—how it was used, and how it could have been used to properly evaluate the riskiness of mortgage loans and the derivatives based on them. The following sources of information will be used in our review:
    a. Demographic data for the U.S. housing market from the Home Mortgage Disclosure Act
    b. Aggregate data on foreclosure rates published in *GH Bank Housing Journa*l (Barth et al.)
    c. The Case-Shiller Home Price Index
    d. A fraud index developed by Interthinx, an ISO business

I. Case Study #1: The Madoff Data

In December 2008, the nation was stunned when Bernard Madoff was arrested for perpetrating one of the world's largest Ponzi schemes. What was remarkable about this fraud was that more than eight years earlier the SEC had been alerted to the fraud by Harry Markopolos. Markopolos, a securities industry executive, testified: "As early as May 2000, I provided evidence to the SEC's Boston Regional Office that should have caused an investigation of Madoff. I resubmitted this evidence with additional support several times between 2000 and 2008."

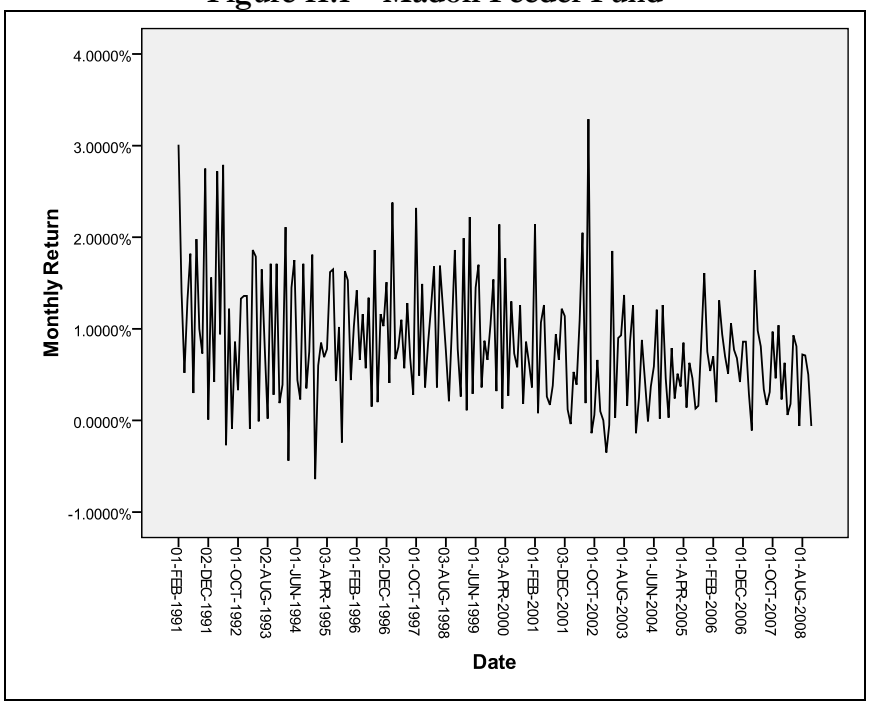
According to ASOP No. 23, "A review of data may not always reveal existing defects. Nevertheless, whether the actuary prepared the data or received the data from others, the actuary should review the data for reasonableness and consistency." The Markopolos testimony, along with

[3] The data is from the hedge funds Fairfield Sentry, Ltd.

the statements of others (Arvedlund 2008), suggest that due diligence was not performed by the "feeder funds" or fund managers, who provided many of the client funds "invested" in the Madoff Ponzi scheme. It also appears that the SEC was lax and did not follow through on what now appears to be obvious clues to the fraud.

Much of our analysis of data will be motivated by the analyses described in the Markopolos testimony (Markopolos 2009). The analysis will be applied to published returns from one Madoff feeder fund. The fund was sold by the Fairfield Greenwich Group. The information was downloaded from the Internet in early 2009. The data is the monthly return for the fund from January 1991 through October 2008. **Figure II.1** displays a graph of the monthly return for this fund.

Figure II.1 – Madoff Feeder Fund



In the testimony, Markopolos described a number of approaches he used to determine that Madoff return data was fraudulent. According to Markopolos, the returns on Madoff's investments were too good to be true. He believed that the "split-strike conversion" strategy that Madoff claimed to use would probably not beat T-Bill returns, especially after expenses are factored in.

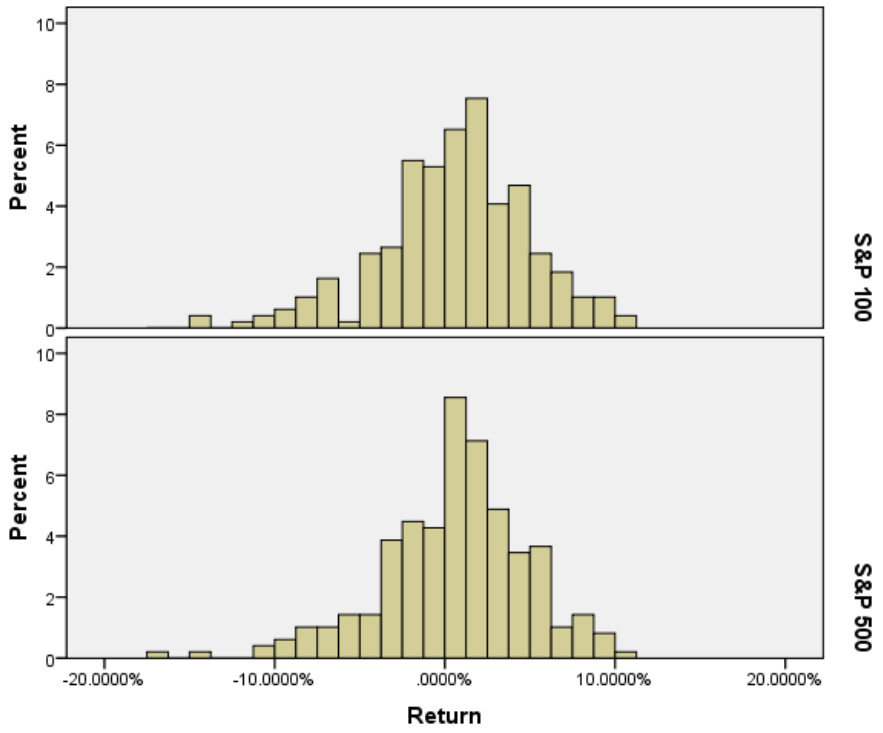
In the analyses in this section we will show that a number of simple graphical and descriptive statistics should have provided red flags or warnings that the Madoff data was fraudulent. These tests include histograms, descriptive statistics and a scatter plot. We also introduce a statistical test, Benford's Law, which is frequently used to detect fraudulent transactions by forensic accountants.

Madoff claimed to have purchased a basket of 30-35 stocks whose returns closely followed the returns of the S&P 100 index. Because it had fewer stocks, this portfolio could be expected to have higher volatility than the index it is tracking. Absent other aspects of the investing strategy to dampen volatility such as the split-strike strategy, which we will address later, Madoff's returns would be expected to share many characteristics with the S&P 100 index.

Test 1: Histogram of Returns[4]

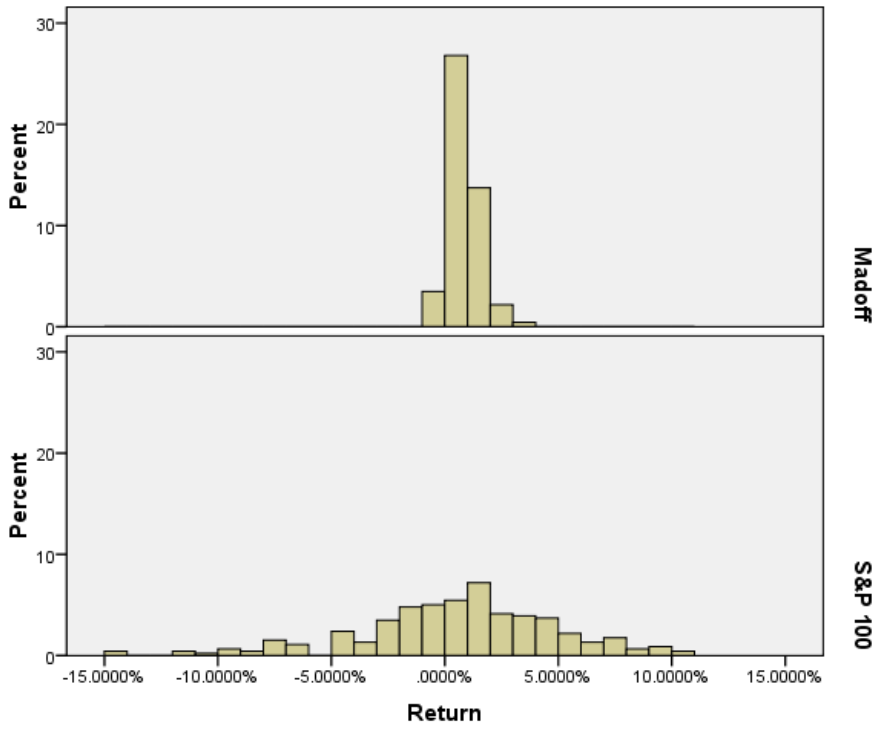
As **Figure II.2** shows, the histograms of the returns for the S&P 500 and its subset of 100 stocks ("S&P 100") are very similar and with an almost bell-shaped distribution. Now compare the histogram of the returns for the S&P 100 and the Madoff Feeder Fund, which was composed of the largest 30-35 of S&P 100 companies (see **Figure II.3**). The Madoff and S&P 100 histograms are dramatically different. The Madoff data is much less dispersed, i.e., it has a much higher peak and much shorter tail than the S&P 100 returns, and has virtually no left tail, suggesting the two are from very different distributions.

Figure II.2 – Histogram of S&P 500 and S&P 100



[4] For Figures II.2 and II.3, the x-axis is the monthly return for the fund shown and the y-axis is the percentage of months with said return. Bins are computed by SPSS software using an automatic statistical rule.

Figure II.3 – Histogram of S&P 100 and Madoff Feeder Fund



Test 2: Descriptive Statistics

Another test Markopolos performed was an examination of simple statistics for a Markopolos fund. We perform a similar analysis below. For comparative purposes, we have provided returns of several stock indices and large mutual funds including:

- The S&P 500. Because it contains 500 stocks instead of the 30-35 in Madoff's fund, it should have lower volatility (standard deviation) than the Madoff data.
- The S&P 100. This is the index Madoff claimed to track. Because it contains 100 stocks instead of the 30-35 in Madoff's fund, it should have lower volatility (standard deviation) than the Madoff data.
- A Balanced Fund that contains a mixture of equities and income producing investments and could be expected to have lower volatility than an equity index (S&P 100). A motivation for investing in a balanced fund is to reduce exposure to risk. A balanced fund is expected to be significantly less volatile and to have somewhat fewer extreme values than a stock-based fund.
- A long-term bond fund that should have the lowest volatility of all the assets as it is composed entirely of bonds.

In producing the following tables, data were limited to returns subsequent to June 1996 because the return series from our comparative mutual funds begins in mid-1996. A surprising result is that the Madoff Feeder Fund has a lower standard deviation than even the bond fund (see **Table II.1**). Indeed, its standard deviation is about 25% of that of the next most volatile asset category. It can also be noted that the Madoff data has a relatively large positive skewness while all the other assets have negative skewness.

Table II.1 – Return Statistics for Different Assets

Asset	Mean	Std. Deviation	Skewness	Kurtosis	N
Balanced	0.46%	2.84%	(0.89)	1.69	149
Long Bond	0.60%	2.40%	(0.36)	2.24	149
S&P 100	0.31%	4.77%	(0.47)	0.47	149
S&P 500	0.30%	4.61%	(0.70)	1.02	149
Madoff	**0.75%**	**0.62%**	**1.01**	**1.25**	**149**

Another statistic that Markopolos commented on was the small number of negative return months for Madoff investments. **Table II.2** displays the negative return statistics for each of the assets. The Madoff fund has a far lower percentage of negative returns than any of the other assets, including a bond fund. According to Markopolos, the probability of such a low percentage of negative return months with a real invested asset is virtually nil.

Table II.2 – Percent of Months With Negative Returns

Asset	Percent of Months
Balanced	39%
Long Bond	37%
S&P 100	44%
S&P 500	42%
Madoff	**6%**
Total	33%

Another point that Markopolos made was that, given a portfolio of 30 to 35 stocks, at least one stock would experience a significant loss[5] during at least one month that would result in a negative return of at least 3% for the portfolio. We see in **Table II.3** that the minimum return for the Madoff fund was a negative 0.6%, well above that of any of the other assets. For the entire 18 years

[5] Significant loss means the stock's value approaches zero.

(as opposed to the 12 years in Table II.2 and II.3[6]) of the Madoff data, the Madoff minimum return was -0.64% versus -14.6% for the S&P 100 for the same period.

Table II.3 – Investment Returns for Different Assets[7]

Asset	Median	Minimum	Maximum
Balanced	0.8%	-11.6%	5.7%
Long Bond	0.9%	-8.7%	11.4%
S&P 100	1.0%	-14.6%	10.8%
Madoff	**0.7%**	**-0.6%**	**3.3%**

Test 3: Benford's Law

Benford's Law is a little known statistical procedure that is used to detect accounting fraud. It is particularly useful in detecting "fake" data. The test is based on the distribution of the first digits of numbers. For instance, Triola (2002) notes that the first digits of amounts on checks tend to follow Benford's law. **Table II.4** below displays the theoretical distribution.

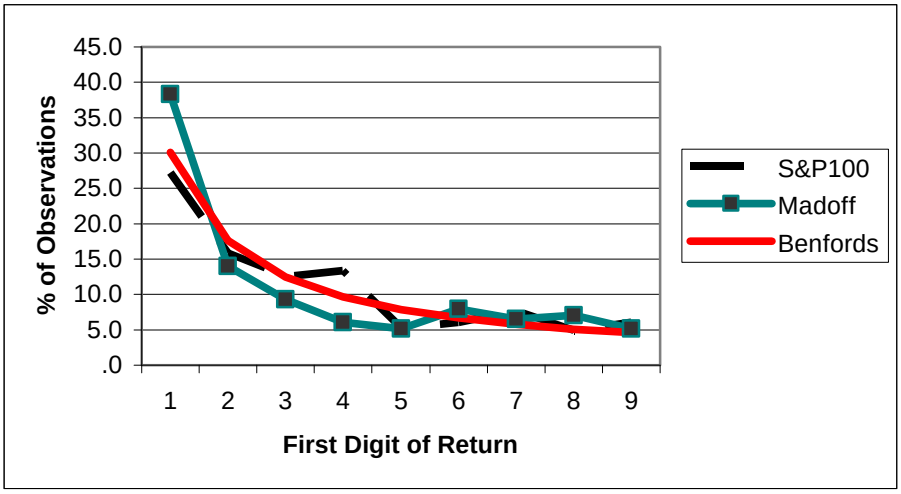
Table II.4 – Benford's Law

First Digit	Proportion
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

A common mistake of people committing fraud is to assume that digits are uniformly distributed. Thus, the perpetrators of fake data tend to fabricate returns whose first digits fluctuate randomly around a discrete uniform distribution. **Figure II.4** displays the distribution of the first digit of the Madoff Feeder Fund's returns compared to the theoretical Benford's Law distribution. For comparison purposes, the distribution of S&P 100 returns are displayed, as we assume these represent the distribution of digits of a "true" random return series.

[6] As noted above, some of our mutual fund data used in the comparisons began in 1996, therefore for these tables we used the same time periods from the Madoff data, thus excluding Madoff returns from 1991 through June 1996.
[7] Percents shown are returns as a percent of assets.

Figure II.4 – Distribution of First Digit of Return



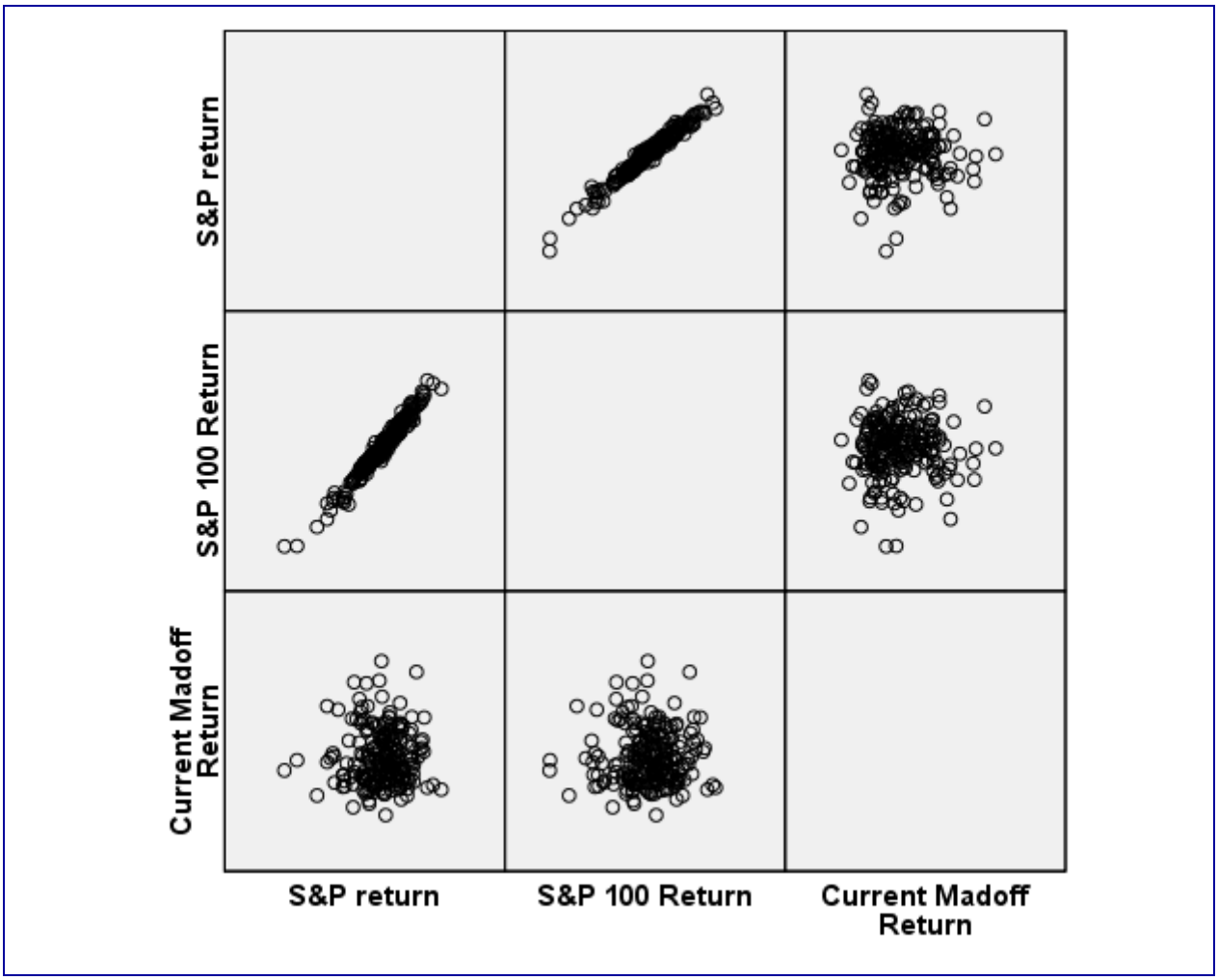
Note that the most obvious discrepancy between the Madoff data and Benford's Law is that the Madoff data displays an excess occurrence of the number 1. The well-known Chi-Square test can be used to compare the actual and expected frequencies of the digits to assess the overall significance of departures of actual from expected.[8] When this test was applied to the Madoff fund data the difference between actual and theoretical frequencies was not significant at the 5% level, although the p-value of 5.9% was close to significant. For comparison, when the Chi-squared test was applied to the S&P 100 data, its p-value was 13.5%. Other researchers have compared the Madoff data to the Benford's Law expectations and have been surprised by the findings. Kedroski opined, "It is interesting to see that any fraud here was sufficiently sophisticated such that the proffered performance numbers were credible from a distributional point of view." Thus, a test that is frequently used to detect fraud, in the case of the Madoff scheme, does not appear to provide compelling evidence.

Test 4: Scatter Plot of Returns vs. a Related Return Series

Figure II.5 displays a matrix scatter plot of the S&P 500 return versus that of the S&P 100 along with scatter plots of the Madoff returns versus both of those indices.

[8] Use the well-known formula $X^2 = \sum_k \frac{(Observed - Expected)^2}{Expected}$ with the Chi-Squared distribution and degrees of freedom of k-1

Figure II.5 – Scatter Plot Matrix of S&P 500, S&P 100, and Madoff Returns

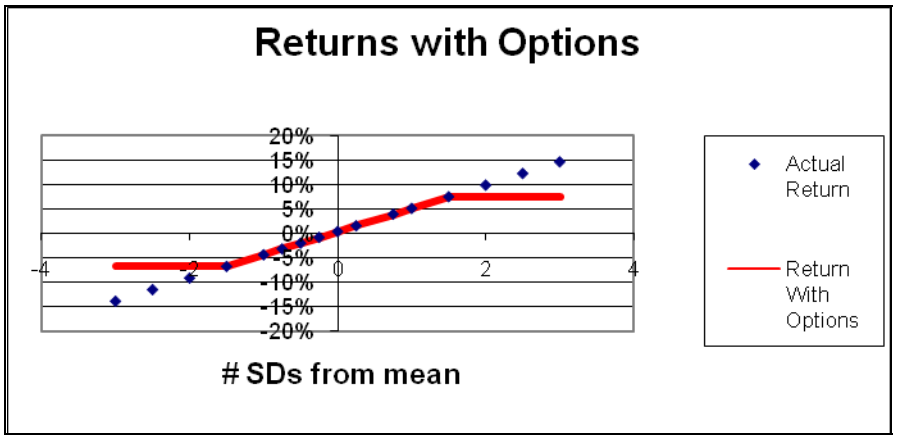


As one would expect, the scatter plot indicates a very high correlation between the S&P 500 and its relative, the S&P 100. However there is no apparent correlation between the S&P 100 and the Madoff fund, even though the Madoff Fund purportedly consists of S&P 100 stocks.

Test 5: The Split-Strike Strategy

In interpreting the graphs and tables, it is necessary to consider the "split-strike" options strategy that Madoff claimed to use. Forray described the strategy and evaluated their likely impact on the descriptive statistics and graphs such as those in this section. As described by Forray (2009), the split-strike strategy involved buying put options to limit downside volatility, while at the same time selling out-of-the-money call options to fund the purchase of the put options. **Figure II.6** is provided to illustrate the split strike options strategy.

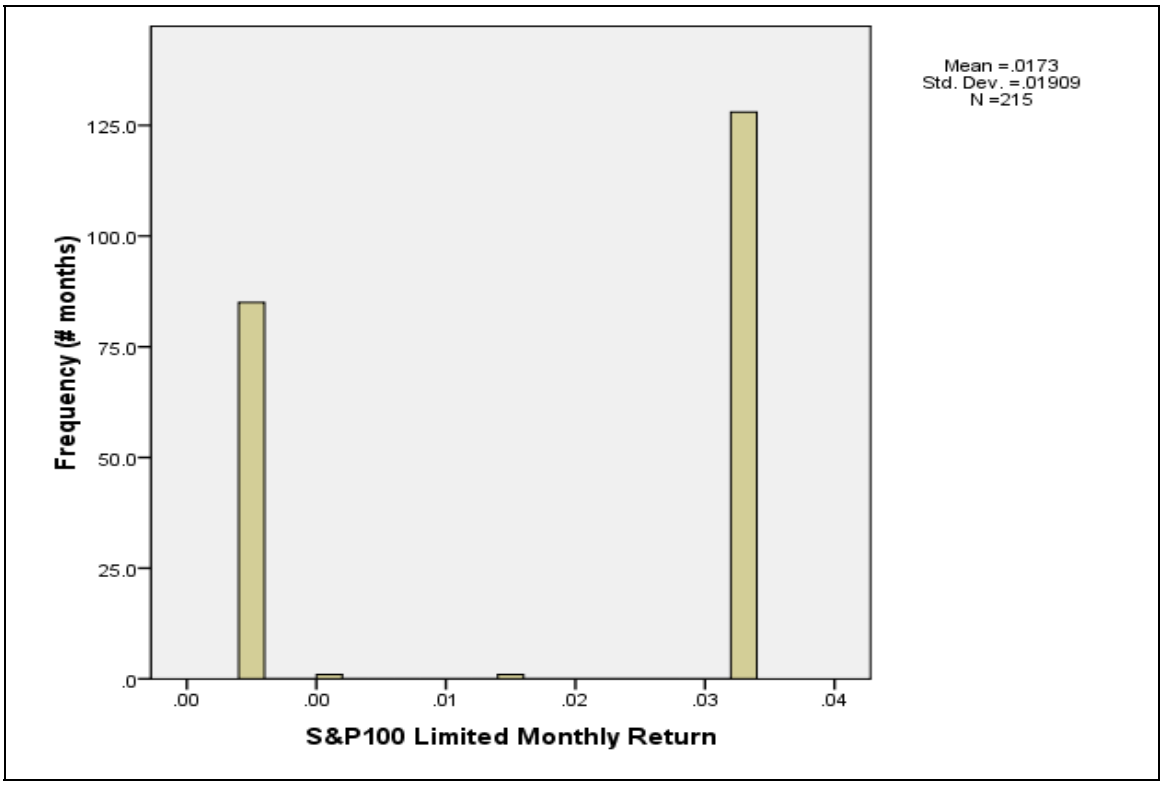
Figure II.6 – Effect of Options on Return



The use of options censors the return of stock funds. The sale of the call options limits the upside potential for the fund. On the down side the purchase of put options limits the downside negative return. For example, if a fund manager buys put options with a strike price of 1.5 standard deviations below the mean (for the S&P 100, approximately -7%), the fund's monthly return will never drop below the strike price (i.e., -7% is the minimum possible return). Alternatively, if returns exceed the strike price of the call options sold by the fund, the funds return is limited by the strike price (say to approximately +7.5%, if the strike price is at +1.5 standard deviations). We could expect the use of the options to reduce the correlations (especially in the tails) between the Madoff Fund and the S&P 100, but that it should still be high. Moreover, transaction costs are incurred to buy and sell options, and these costs depress the returns of the fund. Forray (2009) estimated the net cost of the option transactions to be 0.5% per month, suggesting that the excess return of the Madoff fund compared to the S&P 100 is much more suspect.

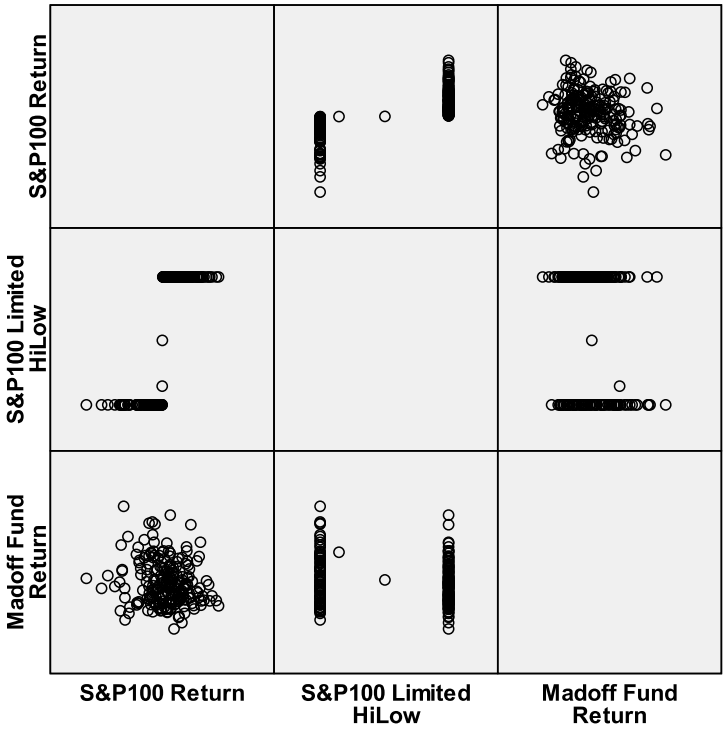
To provide a rough assessment of the impact of limiting returns with options, we capped our S&P 100 returns below at -0.6% and above at +3.3%, approximating the 18-year minimum and maximum in the Madoff fund data. Below we display the histogram of the limited data and note that it is quite different from the histogram in Figure II.3. In particular, the limiting of the returns causes two spikes in the histogram at the lower and upper limit values.

Figure II.7 – Histogram of S&P 100 Returns Limited Above and Below



We also provide a scatter plot matrix displaying the S&P 100, the limited S&P 100 and the Madoff returns. Note that there is still virtually no correlation between the Madoff data and the limited S&P 100 series, while there is still a meaningful correlation (the measured correlation was 0.775) between the S&P 100 and the limited S&P 100.

Figure II.8 – Scatter Plot Matrix, S&P 100, S&P 100 Limited and Madoff Fund



All Tests Recap

We evaluated the Madoff fund data using simple graphical and statistical procedures. We compared the histogram of the Madoff returns to that of a purportedly related return series (if the claim to the split-strike, S&P 100 investment is believed) and found stark differences. We produced a scatter plot of the Madoff returns and found no meaningful correlation, even though there should have been one. We also computed simple descriptive statistics for the Madoff returns and four other comparative return series. The Madoff returns (1) had a much higher mean and (2) a much lower standard deviation than all comparison categories, including a balanced and long term bond fund whose return volatility should be tempered compare to an equity fund. We also found that the Madoff data was positively skewed compared to the negative skewness of the other series.

A more sophisticated test, Benford's Law, was applied to he Madoff returns and failed to provide evidence of fraud. While Benford's Law is a handy tool used by fraud experts to uncover "fake" accounting transactions, it would not likely be known to the typical fund manager screening potential investments for clients. It is possible that Madoff was familiar with the rule and prepared reports to investors that would not trigger an investigation by those applying the rule. Nonetheless,

the dramatic departure of the Madoff returns from reasonably expected statistical patterns should have triggered a high level of concern on the part of fund managers and the SEC.

II. CASE STUDY #2: MORTGAGE DATA

The underwriters of mortgage loans had available to them an abundance of loan level data that could have been used to monitor the overall quality of their book of loans. In addition to their internal data, a number of external data sources were available. One of the popular external databases was LoanPerformance.[9] This database is cited frequently in the financial crisis literature (for instance, Demyanyk and Hemert 2008). As it is a commercial database we do not use it in our analysis, but will cite the results of others who had access to it.

Another database that has been used to monitor the performance of the mortgage industry is the Home Mortgage Disclosure Act (HMDA) database. Enacted by Congress in 1975 and implemented by the Federal Reserve Board's Regulation C, the HMDA requires lending institutions to report public loan data. This data is publically available from a U.S. government Web site.[10] The authors analyzed publicly available HMDA data for loans originating in 2007.[11] As we were limited in the number of years that were available, we chose a cross-sectional comparison, though, as we will show in the next section, time series statistics from another source portrays the unfolding of the crisis over time.

Our cross-sectional comparison uses data for six states – Alabama, Arizona, Florida, Michigan, Nebraska, and Nevada. According to Realty Trac® in its May 2009 U.S. Foreclosure Market Report™, released June 11, 2009,[12] the relative rank of foreclosures (shown in parentheses) for these six states are shown in **Table III.1**.

Table III.1

High Rates of Foreclosure	Low Rates of Foreclosure
Arizona (4)	Alabama (30)
Florida (3)	Nebraska (45)
Michigan (6)	
Nevada (1)	

[9] See www.loanperformance.com for information about their data.
[10] Currently it can be obtained from http://www.ffiec.gov/hmda/.
[11] Data for other years was not available in data set format from the HMDA site.
[12] http://www.realtytrac.com/gateway_co.asp?accnt=137300

Variables in the Data

There are 36 variables in the HMDA data. Most of these are related to the nature of the loan, but geographic and demographic information is also available. **Table III.2** is a list of some variables that could be useful in assessing the riskiness of loans. Some of the variables, such as census tract number, could be useful in incorporating external data not contained in the HMDA database.

Table III.2

Variables in HMDA Data
Agency Code
Loan Purpose
Loan Type
Property Type
Occupancy
Loan Amount(000s)
Preapproval
Action Type
County
Census Tract Number
Applicant Income (000s)
Purchaser_Type
Denial Reason
Rate Spread
Lien Status
Rate Spread
Lien Status

Table III.3 provides descriptive statistics for three of the variables that should be useful in assessing the riskiness of loans – loan amount, applicant's income, and rate spread (spread of loan interest rate to index interest rate, often LIBOR[13] or other internationally recognized benchmark). Note the heavy tails and skewness of the distributions, reflecting the presence of extreme values for the variables. Note also that some extreme values (i.e., loan amounts of $99.999 million) were assumed to be suspect and were not included in the analysis.

[13] LIBOR is the London Interbank Offered Rate. LIBOR is an interest rate at which banks can borrow funds, in marketable size, from other banks in the London interbank market. LIBOR is fixed on a daily basis by the British Bankers' Association.

Table III.3 – Descriptive Statistics for HMDA Florida Data[14]

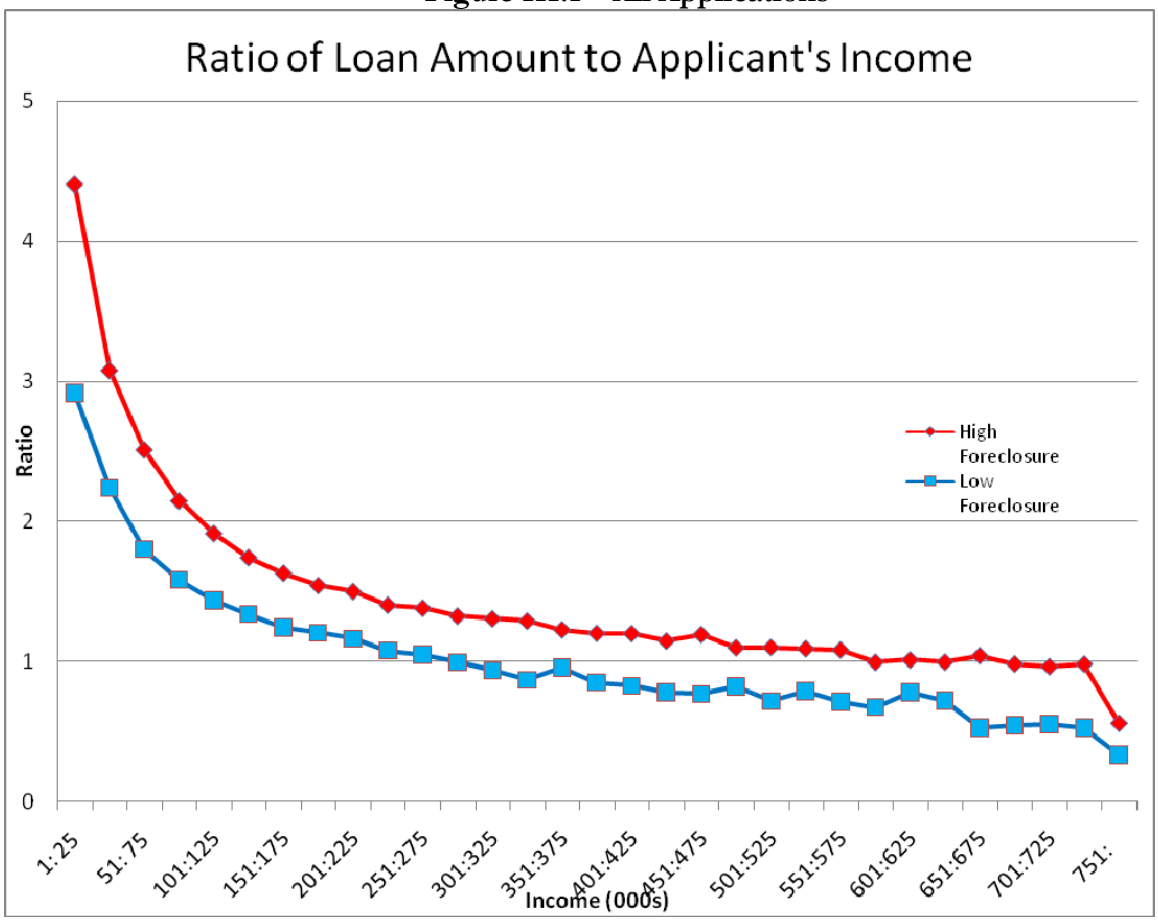
		Loan_Amount_000s	Applicant_Income_000s	Ratespread
N	Valid	1773450	1773450	159203
	Missing	0	0	1614247
Mean		206.52	114.20	5.0495
Median		171.00	75.00	4.7400
Skewness		18.549	16.011	.827
Std. Error of Skewness		.002	.002	.006
Kurtosis		1817.752	473.308	.775
Std. Error of Kurtosis		.004	.004	.012
Minimum		2	2	3.00
Maximum		45500	9981	30.36
Percentiles	5	31.00	28.00	3.0800
	10	50.00	35.00	3.1700
	20	90.00	45.00	3.3800
	30	120.00	54.00	3.6800
	40	147.00	64.00	4.0900
	50	171.00	75.00	4.7400
	60	198.00	88.00	5.4100
	70	229.00	105.00	5.9800
	80	275.00	136.00	6.5600
	90	364.00	204.00	7.3600
	95	468.00	300.00	8.0500

The loan and income information were used to compute a loan-to-value ratio, as loan amounts that were excessive relative to the income of the applicants were considered one of the key risk factors in the mortgage crisis.

Figure III.1 shows the ratio of Loan Amount to Applicant's Income for the six states with the data aggregated into two groups – high-foreclosure states versus low-foreclosure states. For this analysis, we have removed records from the database that have no income reported. The states with a higher incidence of foreclosures show a consistently higher ratio of Loan Amount to Applicant's Income across all income buckets.

[14] For our analysis for rate spread in this section, the percentiles and other descriptive statistics ignored records with missing values, i.e., coded as "NA."

Figure III.1 – All Applications



The logical question is: Was the higher riskiness of the loans based on the loan-to-income ratio reflected in the rate spread[15] for the loans? While **Table III.4** does not show a dramatic difference in the distribution of loan applications by rate spread, the states with a higher foreclosure rate have a slightly higher percentage of loan applications with a zero basis point difference. One would expect the opposite to be the case. Additional examination of the data showed that non-owner occupied houses and refinanced mortgages, as opposed to original mortgages, had lower rate spreads on average.

A question that arises is: does this reflect a major problem with the data reported into the HMDA database? That is, we observed that 90% of the records may have been incomplete with respect to this field, as the rate spread value was reported as "NA." In the computations underlying Table III.4, the NAs were treated as if the rate spread was zero.

[15] Spread of loan interest rate to index interest rate, often LIBOR or other internationally recognized benchmark.

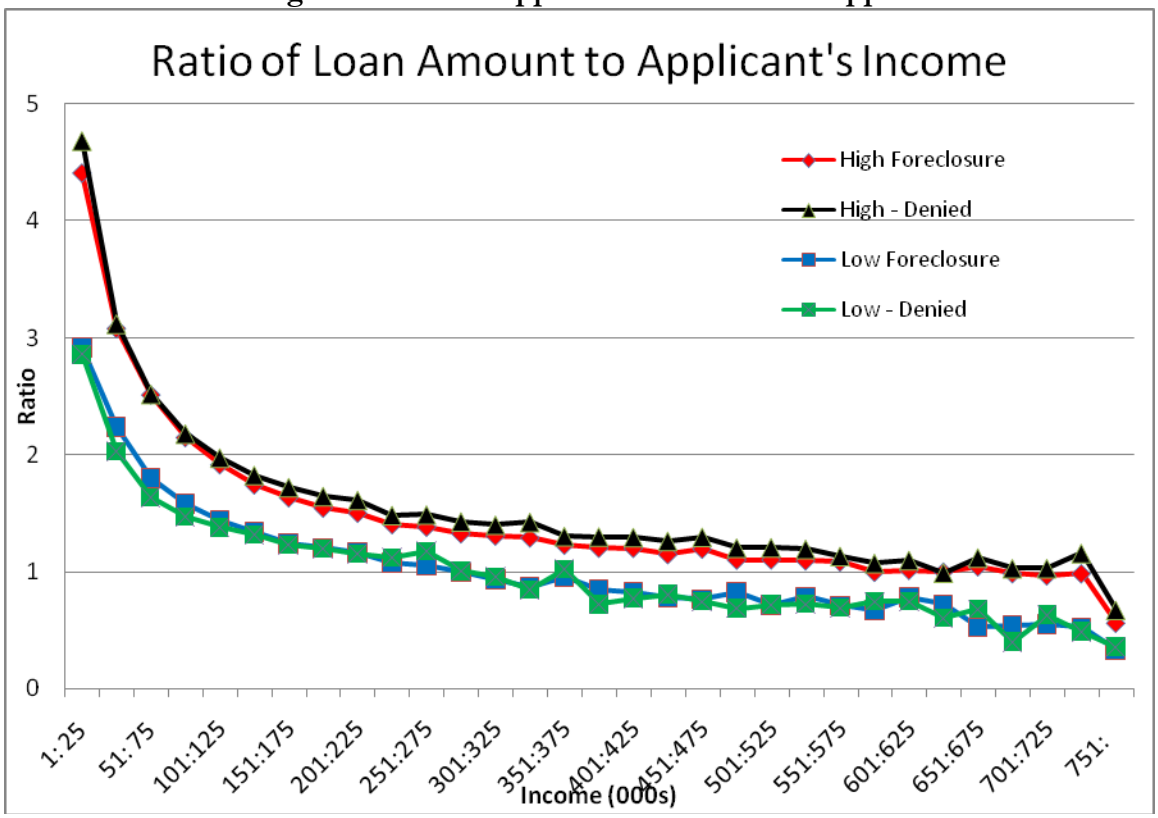
Table III.4

	Loan-to-Income Ratio		Distr of Rate Spread	
Rate Spread	Low Foreclosure States	High Foreclosure States	Low Foreclosure States	High Foreclosure States
None[16]	1.56	1.82	90.0%	91.7%
3.00: 3.99	1.51	1.92	3.2%	3.1%
4.00: 4.99	1.47	1.95	1.9%	1.2%
5.00: 5.99	1.22	1.56	1.9%	1.5%
6.00: 6.99	1.36	1.73	1.3%	1.3%
7.00: 7.99	1.26	1.33	0.9%	0.7%
8.00: 8.99	0.71	0.65	0.3%	0.3%
9.00: 9.99	0.50	0.46	0.2%	0.1%
10.00+	0.41	0.39	0.2%	0.1%
Total	1.54	1.81	100.0%	100.0%

Figure III.2 shows the ratio of loan amount to applicant's income for the six states again but showing all loan applications versus those applications that were denied for the two groups – high-foreclosure states versus low-foreclosure states. In this case, we see no discernable difference in this ratio between all applications and those that were denied.

[16] In this Table, records with a rate spread of NA were assumed to be zero.

Figure III.2 – All Applications vs. Denied Applications



Time Series Patterns: Demyanyk and Helmert (2008) Analysis Using Loan Level Data

Evolution of some of the key risk variables, such as loan-to-value ratios, can provide valuable insights into the riskiness of the loan portfolio and how it changed over time. Using data unavailable to us, Demyanyk and Helmert performed such an analysis, based on a history of loan level data from LoanPerformance.com. **Figure III.3** displays a time series of the loan-to-value ratio and shows a decline in loan-to-value ratio in 2002, followed by a steady increase through 2006.

Figure III.3 – Time Series of Loan-to-Value

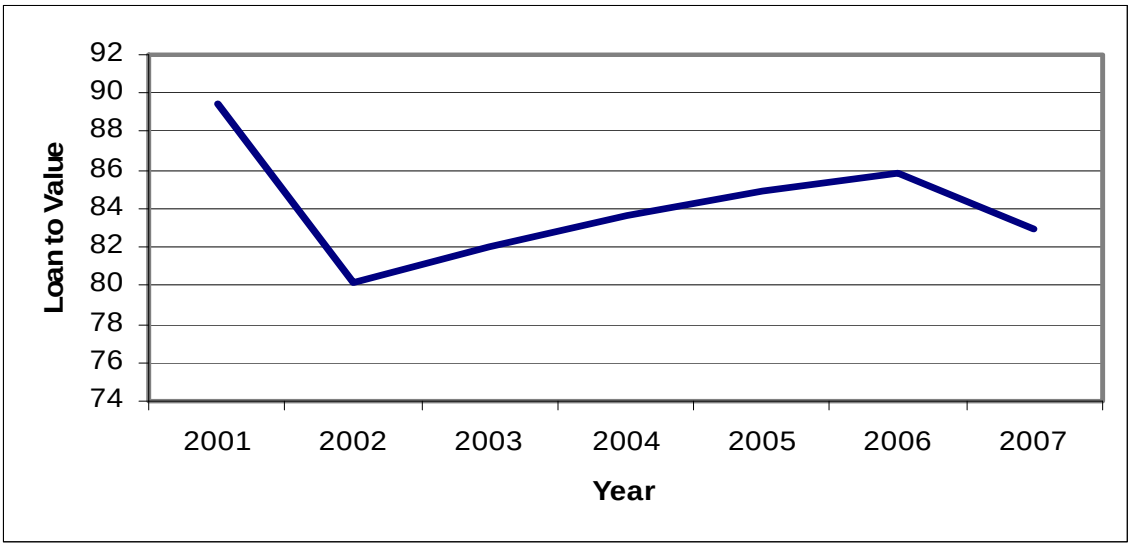


Figure III.4 displays the number and size of subprime loans from 2001 to 2007. It is clear that both the volume and size of such of these loans increased at extraordinary rates. Through 2005 the number of subprime loans increased more than five times and the dollar-value of subprime loans increased by almost a factor of eight. It should be noted overly aggressive growth strategies are a key factor in the development financial bubbles (Ferris, 2009). In addition, an analysis by Carson and Dastrup (2009) indicated that the large percentage of subprime loans (compared to total loans) was a key factor in the mortgage crisis.

Figure III.4 – Subprime Loans

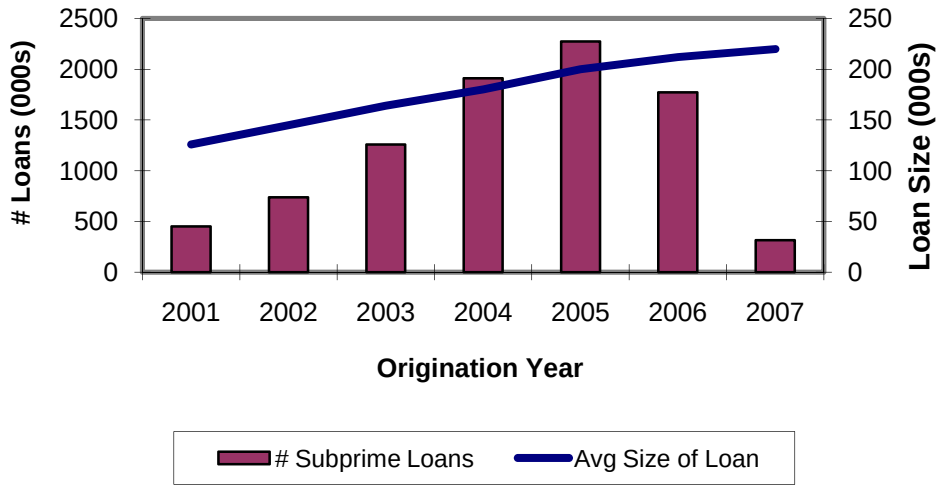


Figure III.5 – Balloon Payments and Documentation Percent

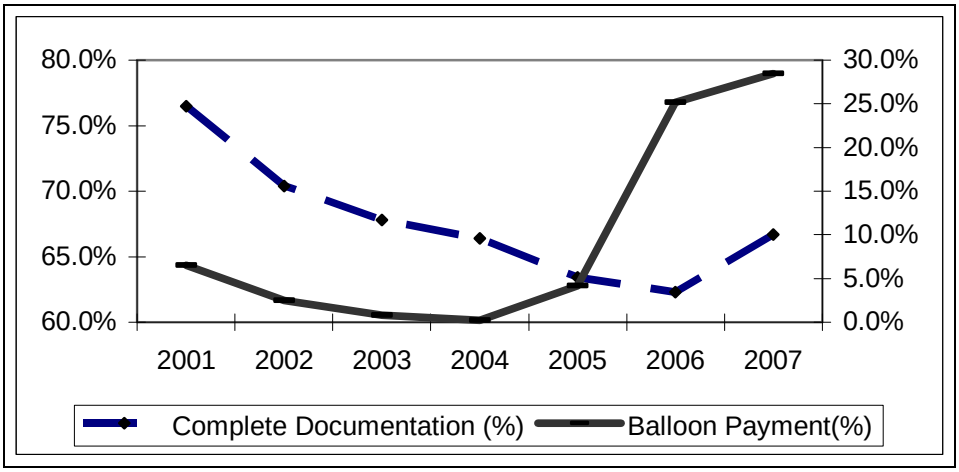


Figure III.5 presents the percentage of files with complete documentation and the percent with balloon payments. The graph indicates that over time documentation standards deteriorated. This is consistent with the appearance of categories of loans referred to as "liar loans" or NINJA (no income, no job, no assets) loans. In addition, Figure III.5 indicates that the percentage of loans with balloon payments ballooned as the mortgage crisis reached its peak. Such loans require the borrower to make a balloon payment upon maturity of the loan, or in the case of some subprime loans, several years into the loan. Without an infusion of income such as from an inheritance or from a litigation settlement, such funding could be hard to produce. In the case of subprime loans, this might force the borrower into another round of refinancing.[17]

Taken together, these descriptive statistics indicate that simple descriptive statistics derived from loan level data indicated the development of significant risks in loan portfolios. When reviewed over time, the statistics indicated deterioration in the quality of loans. When the data was viewed from a cross-sectional perspective, the statistics differentiated problematic from non-problematic states.

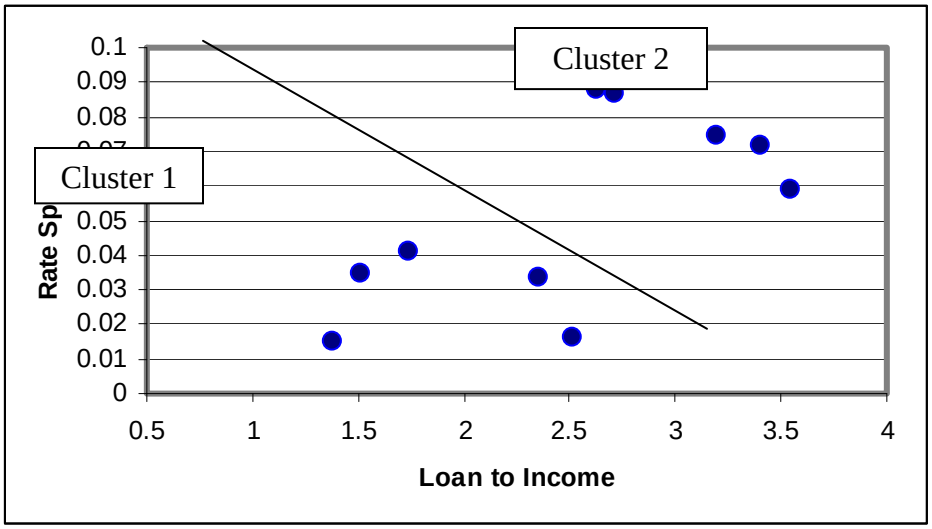
Can Loan Level Data be used to Model Foreclosures?

Can the loan level data be used to model foreclosures as an aid to underwriting loans? Though limited in the number of potential predictive variables compared to a commercial loan database, the HMDA data can be used to illustrate modeling techniques that can be applied to model foreclosures.

[17] The literature suggests that was the intent but foreclosure was also a possible outcome.

Since that data has no dependent variable, that is, we do not know which loans have already nor will default subsequent to being underwritten, we illustrate the unsupervised learning technique of clustering. Clustering can be used to group records, in this case, loan applicants with similar characteristics together. A simple illustration using hypothetical simulated data is supplied in **Figure III.6**. Suppose that data consisted of two loan characteristics. When we graph the values for the two variables the records seem to fall into two groups (separated on the graph by a line), one with low values on both variables and another with high values on both variables. One of these groups, also referred to as clusters, may be correlated to a variable of interest to us such as propensity to default. The clustering methodology applies a statistical procedure that maximizes the differences between clusters on values of the clustering variables. Some variables will inevitably be more significant in distinguishing groups from each other. The methodology used to cluster data is introduced by Francis (Francis 2006) and Sanche (Sanche 2006) and is widely described in statistics text books and is outside the scope of this paper.

Figure III.6 – Loan-to-Income Clusters



In order to perform clustering, data from the HMDA database for Florida were aggregated to the ZIP code level. The loan's census tract number was used to associate a ZIP code with each loan application record in the HMDA database. Then ZIP code level statistics were computed from predictive variables such as income, loan amount, and loan-to-value ratio.[18] These statistics included means, medians, and the percent of loans exceeding a high threshold (such as the 90th percentile value for the variable). For assessment of the effectiveness of clusters in grouping high-risk versus

[18] Data used in this analysis were from the state of Florida.

low-risk loans, data from an external source was incorporated into the data. The data is from LISC Research.[19] A model was used to score each ZIP code on the extent of its foreclosure, delinquency, and subprime problem to assist in identifying areas of greatest need for stabilization. In developing the score, information from the Mortgage Bankers Association Delinquency Survey and reports from McDash Analytics, a vendor of loan information, as well as other sources, was used. Note that a limitation of the data is that the foreclosure scores are a calendar period statistic while the HMDA data is organized by origination year. We believe that this mismatch will tend to temper the impact of relationships between HMDA loan characteristics and the LISC scores.

[19] Obtained from www.housingpolicy.org.

Table III.5 displays the mean of the ZIP code level characteristics for each of the three clusters assigned by the clustering procedure.

Table III.5 – Means On Variables[20]

	Cluster		
	1	**2**	**3**
Avg Loan Amount	297.23	566.96	163.80
Average Income	165.71	356.66	87.26
Mean LTV[21] Ratio	2.53	2.38	2.48
Rate Spread - mean	4.84	4.54	5.05
Median LTV Ratio	2.29	2.09	2.31
Median Rate Spread	4.40	3.95	4.67
Percent Applicants High LTV	4.4	3.8	4.5
Pct Applicants High Rate Spread	4.7	4.5	5.6
Percent Manufactured, Multi Family Houses	1.9	.4	6.1
Pct Home Improvement	57.8	56.5	65.6
Percent Refinance	52.4	52.5	57.3
Pct Owner Occupied	18.1	28.4	13.5

It can be seen that the clusters differ significantly on income level, loan amount, and the percentage of homes that are owner-occupied, but less so on loan-to-value and rate spread. Based on the mean statistics for the various loan risk variables, it seems that cluster 2 has the highest incomes and percent owner-occupied, while having the lowest loan-to-value ratio and rate spreads. It might be expected this cluster would have the fewest problem loans. Cluster 1 has intermediate values on income, loan amount, rate spread, median (but not mean) loan-to-value ratio, and the percent owner-occupied. This suggests these ZIP codes may have a less severe, but non-negligible problem than cluster 3.

Table III.6 displays the average score for each cluster for the foreclosure, delinquency and subprime scores. Keeping in mind that a high value indicates a severe problem, it appears that on all three problem-mortgage measures, cluster 2 has the least severe problem, followed by cluster 1, and then cluster 3 has the most severe problem.

While many limitations affect this exercise, we believe it illustrates that loan level data available to lenders during the period when the housing bubble was reaching its extreme could have been used to develop models for use in underwriting and in avoiding high-risk mortgages.

[20] Statistics for rate spread variable based only on non-zero records.
[21] LTV = Loan to value

Table III.6			
Cluster Grouping	**Intrastate Subprime Component Score**	**Intrastate Foreclosure Component Score**	**Intrastate Delinquency Component Score**
1	3.059	2.315	7.325
2	0.751	1.074	2.702
3	7.830	3.119	12.482
Total	6.774	2.920	11.288

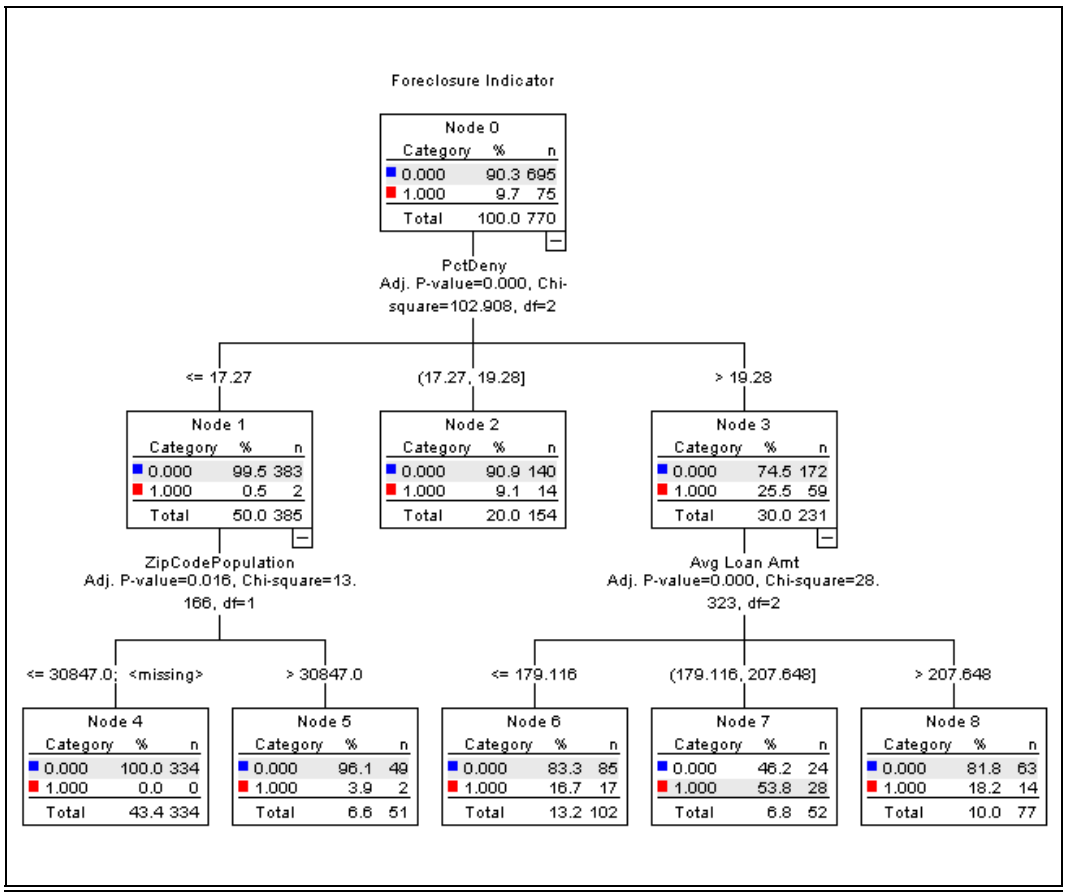
When a dependent variable is present, say loan data dating back to the late 80s and early 90s when the last housing bubble popped, a data mining method can be used to predict foreclosure problems. For illustrative purposes, one data mining method was applied to the ZIP code level data used in the previous section for clustering.[22] For this illustration, the dependent variables are the foreclosure and subprime scores from the LISC data. Of course, the values for these variables would not have been available at the time the mortgages were underwritten, however, we believe that other relevant data not available to us, such as historic default and foreclosure rates, should have been available to mortgage modelers.

The data mining method used was decision trees. This method allows the analyst to quickly search for important relationships in the data. The procedure allows the incorporation into the model of complex relationships, such as interactions and nonlinearities. In addition, many of the software tools used for tree modeling rank the predictive variables in importance. Derrig and Francis (2008) introduce a number of the tree procedures and we recommend this as well as DeVille (2006) to the interested reader seeking background information about trees.

The result of applying trees to model foreclosure problems is presented in **Figure III.7**.

[22] Only one state, Florida was used in this analysis. LISC sores are relevant only within a state, not across different states.

Figure III.7 – Tree-Based Foreclosure Model



In **Table III.7** the ranking of variables from a tree modeling procedure is presented. Note that the top two variables in predicting foreclosures are derived from the denial rates in the ZIP code. We suspect that this reflects the realization, partway through the 2007 year, of the extent of the mortgage problem, along with at least a partial return to traditional underwriting standards. The use of the house (single versus multifamily and manufactured) and the rate spread were also important predictors.

Table III.7

Independent Variable	Importance	Normalized Importance
Denial Percent	.027	100.0%
Mean Denial Score	.027	99.9%
PctApprove	.024	88.5%
ZipCodePopulation	.020	72.6%
PctManuMultiFamily	.019	69.5%
Median Rate Spread	.017	61.6%
LoanPurpose	.016	60.5%
HouseholdsPerZipcode	.015	56.1%
Mean LTV Ratio	.014	52.7%

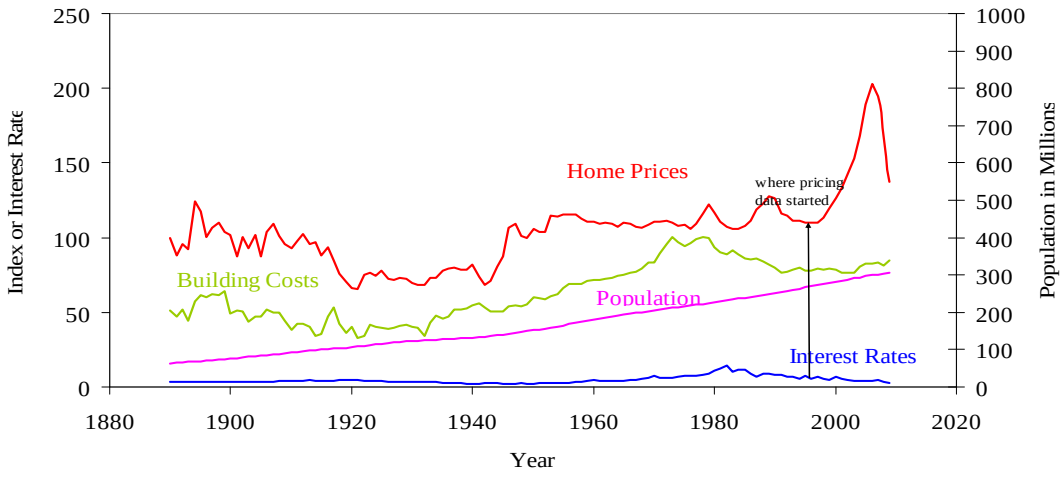
We believe this exercise with the HMDA data justifies the belief that well-known modeling techniques could have been applied to loan level data to predict the loans with a high likelihood of default.

Do Housing Prices Go Down?

A key and virtually universal belief underlying the pricing of many mortgage products and the credit analysis of the products was that housing prices never decline, especially when averaged over large geographic areas, such as the entire U.S. What is most surprising about this belief is that many of us may remember experiencing such a decline in the late 80s and early 90s (depending on where one lived).

A simple check of publically available data would have exposed this fallacy. In **Figure III.8**, we present the widely cited Case-Shiller Home Price index, along with some comparative indices.

Figure III.8 – Housing Price Time Series[23]



Mortgage Fraud: Could We Have Detected the Problem Sooner?

During the recent economic turndown, the incidence of mortgage foreclosures has risen substantially. This has caused an unprecedented drain on the United States economy as the federal government shores up the various financial institutions involved with the mortgages that are now defaulting. But could we have foreseen the escalation in foreclosures? Were there indicators to what was coming? Fraud is believed to have played a significant role in the subprime crisis. In his book *Confessions of a Subprime Lender* (2008), Richard Bittner stated that at the time he exited the business because of his perception of the irrationality of the mortgage market; approximately 70% of the loan applications sent to him contained some level of fraud/misrepresentation.

The authors reviewed Interthinx Fraud Risk Index data, which tracks the risk of mortgage fraud throughout the United States. The Fraud Risk Indices are calculated based on the frequency with which indicators of fraudulent activity are detected in mortgage applications processed by the Interthinx FraudGUARD® system.[24] The Fraud Risk Indices are based on detailed transaction data

[23] Data for Figure III.8 in Robert J. Shiller, *Irrational Exuberance*, 2nd. Edition, Princeton University Press, 2005, Broadway Books 2006, also Subprime Solution, 2008, as updated by author. Graph and data underlying it was obtained from www.iii.org. A publicly available housing price index, the Case-Shiller index, is used in the graph, and is available for download from the Standard and Poor's web site..

[24] FraudGUARD®system is a leading loan-level fraud detection tool available to lenders and investors from Interthinx, an ISO business.

from loan applications supplied by lenders. While the Interthinx FraudGUARD® system provides a check on over 300 items per application, these indices focus on a sub-set of the checks that are most indicative of the type of fraud represented by the indices—property valuation, identity, occupancy, and employment/income. The indices provide a means of comparing geographic regions or time periods—it is the relative values of the indices that are important and not the absolute value of a particular index. The definitions of the fraud risk indices are provided in Appendix A.

For this section of the analysis, we reviewed the indices for Alabama, Arizona, Florida, Michigan, Nebraska, and Nevada, which are the same states we used in the cross-sectional analysis. To develop the index for the "grouped" states (to simplify the presentation), we took the simple average of the state indexes for the states in each group.

While hindsight is always 20-20, the Overall Risk Index (**Figure III.9**) during 2004 may have been a leading indicator of the trouble that was brewing in the mortgage industry. The four states with high foreclosure rates in 2009 have consistently had an index over 100. The index is also rising again in Nevada. In contrast the index for Alabama and Nebraska were consistently lower. **Figure III.10**, the Property Value Fraud Risk Index, shows the same pattern of high risk for fraud during 2004.

Figure III.9 – Overall Fraud Risk Index

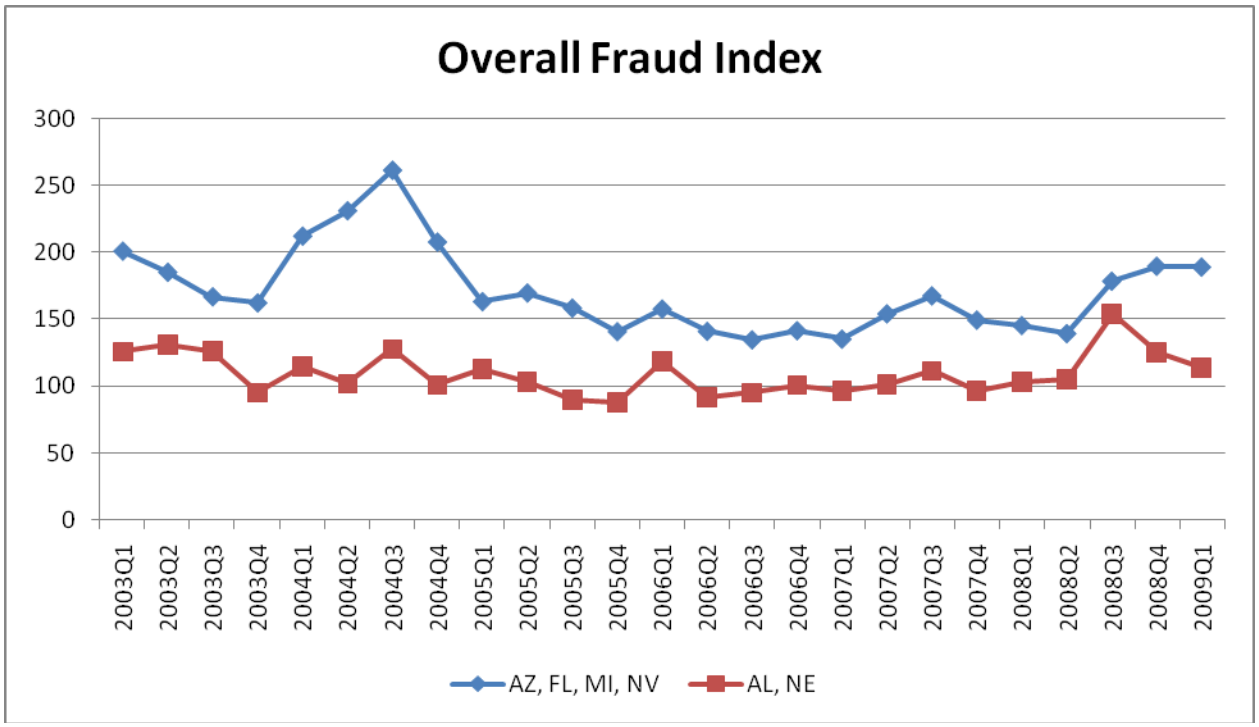
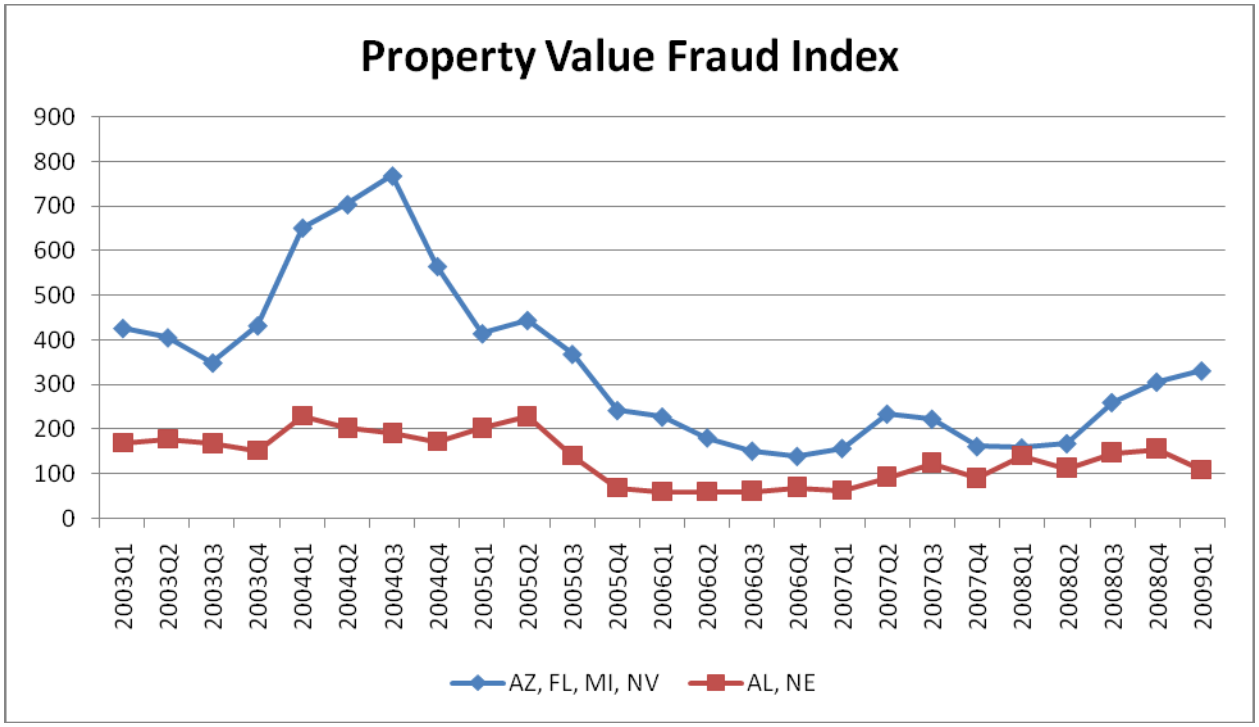


Figure III.10 – Property Value Fraud Risk Index



Mortgage Data Conclusions

ASOP No. 23 prescribes that actuaries use data that are complete and that are appropriate for the analysis. In ignoring the information prior to 1994 that housing prices declined historically on several occasions, the credit rating agencies used data that was incomplete to the point of absuridity. The analysts also failed to consider the limitations of their data (and perhaps to adjust subjectively for it) and to disclose those limitations.

The foregoing analyses also leads the authors to believe that:

- Simple descriptive statistics and predictive models may have helped avoid the crisis that developed in late 2007.

- Monitoring the make-up and quality of the loan portfolio is imperative and provides valuable insights into the riskiness of the portfolio. This is a continuous process as loans are underwritten and added to the portfolio. This is analgous to property/casualty insurance companies monitoring their probable maximum loss (PML) as they underwrite coastal risks in the United States.

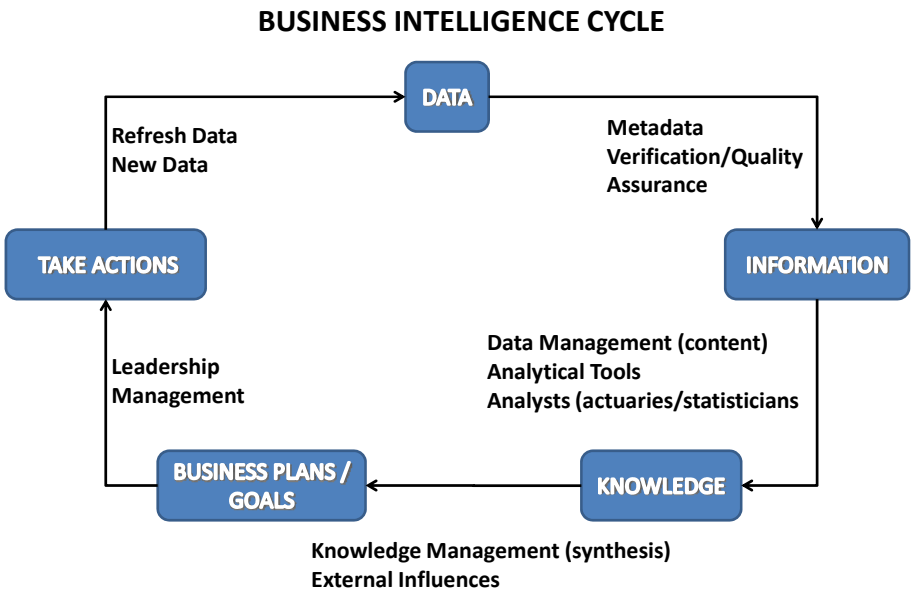
- Loan applications should be reviewed for not only quality of the risk, but also for completeness and accuracy of the application to judge its risk quality.

III. CONCLUSION: COULD BUSINESS INTELLIGENCE HAVE HELPED AVOID THE CRISIS?

Howard Dresner of The Gartner Group defined business intelligence as "a set of concepts and methodologies to improve decision making in business through the use of facts and fact-based systems." The authors believe that the signals were there in the data if the investors and those involved in the mortgage process had done the following:

- Developed systems to collect quality data
- Routinely reviewed simple descriptive statistics from such loan data as were available
- Used publicly available aggregate statistics to validate crucial model assumptions
- Used various analytical tools to convert the data into information to be used in the business process
- Applied leadership and businss aucumen to act on the analytics

Business intelligence coupled with the data management concepts espoused in the recent paper *Actuarial I.Q. (Information Quality)* may help us avert these types of problems in the future. We believe that the financial crisis is the poster child for the failure to apply these business intelligence concepts as visualized in the following figure.



In the case of the Bernard Madoff fraud, a simple review of published return data would have uncovered the unreasonableness of the data. In the case of the mortgage crisis, a review of loan level statistics should have provided an early warning that the quality of mortgages was deteriorating. Moreover, publicly available data would have indicated that housing prices have declined a number of times in the past and that it was extremely dangerous to use models that assume housing prices, in the aggregate, never decline.

Would more training of the principals involved in the financial crisis on data quality and business intelligence have prevented some of the damage? Was the crisis a result of ignorance, or were the financial incentive so overwhelming that business fundamentals were ignored? News sources indicated that inexperienced people were given significant responsibilities, both in the Madoff case (Scheer 2009) and and in pricing mortgage products (Bestiani 2009). Nevertheless, according to Bloomberg Markets (Helyar et al. 2009), some fund managers provided client funds to Madoff, even though they suspected that Madoff was involved in a kind of fraud referred to as "front-running."[25] The implication is that they were willing to go along with an illegal scheme if it brought extra returns to their funds and if they thought the risk to themselves and their invesors, should the fraud be detected, was low.

[25] Front-running is a type of insider trading where brokers trade ahead of the markets.

There are also indications that some data used in pricing and underwriting mortgages was intentionally ignored. Muolo et al. (2008) cite an executive from the compliance firm Clayton Holdings, who informed a rating agency about a number of "exceptions" (i.e., from underwriting standards) in a loan protfolio and was told, "We do not want to see it. It does not fit our model."[26] Bestiani (2009) notes that the rating firms with the most optimistic evaluations were given the rating business by lenders and investment firms. In a November 2, 2008 article titled "Was There a Loan it Did Not Like?" *New York Times* reporter Morgenson describes the travails of a senior underwriter at Washington Mutual who at the height of the bubble was pressured to approve loans that she felt were obviously flawed, and in some cases blatantly fraudulent. At this point, we are not privy to the motivations of those who contributed to the financial crisis. Nonetheless, both inexperience and "the moral hazard problem" appear to have contributed to the poor use of data and failure to apply the concepts of business intelligence.

[26] Muolo et al., 2008, , p. 283.

APPENDIX A: INTERTHINX FRAUD INDEX DEFINITIONS[27]

The Interthinx Fraud Risk Indices consist of the Mortgage Fraud Risk Index, which measures the overall risk of mortgage fraud, and the property valuation, identity, occupancy and employment/income indices, which measure the risk of these specific types of fraudulent activity.

The Mortgage Fraud Risk Index considers 40+ indicators of fraudulent activity including property misvaluation; identity, occupancy and employment/income misrepresentation; non-arms-length transactions; property flipping; straw-buyers; "silent seconds;" and concurrent closing schemes. The four type-specific indices are based on the subset of indicators that are relevant to each type of fraudulent activity.

Each index is calibrated so that a value of 100 represents a nominal level of fraud, a value calculated from the occurrence of fraudulent indicators between 2003 and 2007 in states with low foreclosure levels. For all five indices, a high-value indicates an elevated risk of mortgage fraud and each index is linear to simplify comparison across time and location.

The Interthinx Indices are leading indicators based predominantly on the analysis of current loan originations. FBI and FinCEN reports are lagging indicators because they are derived primarily from suspicious activity reports (SARs), the majority of which are filed after the loans have closed. The time lag between origination and the SAR can be several years. For this reason, the Interthinx Fraud Risk Indices' top fraud geographies and type-specific findings may differ from FBI and FinCEN fraud reports.

[27] *Mortgage Fraud Risk Report*, Interthinx, Inc, August, 2009.

IV. REFERENCES

[1.] Arvedlund, Erin, "Innocence Lost," Portfolio.com, December 17, 2008, http://www.portfolio.com/news-markets/top-5/2008/12/17/madoff-barrons/.

[2.] Barth J., et al., "A Short History of the Subprime Mortgage Market Meltdown," *GH Bank Housing Journal,* Milken Institute, January 2008, http://www.ghb.co.th/en/Journal/Vol2/04.pdf.

[3.] Bestiani, R., "The Perfect Financial Storm," Working Paper #44, 2009.

[4.] Bitner, R., *Confessions of a Subprime Lender*, Wiley, 2008.

[5.] Carson, R. and S. Dastrup, "The Housing Price Bubble: Data Mining to Explain The Housing Price Fall Across Metropolitan Areas," presented at Salford Data Mining Conference, August, 2009.

[6.] CAS Committee on Management Data and Information, "Data Quality White Paper," CAS *Forum,* Winter 1997, pp. 145-168, http://www.casact.org/pubs/forum/97wforum/97wf145.pdf.

[7.] CAS Data Management Educational Materials Working Party, "Actuarial I.Q. (Information Quality)," CAS *Forum*, Winter 2008, pp 136-195, http://www.casact.org/pubs/forum/08wforum/actuarialIQ.pdf.

[8.] Demyanyk, Y. and O. Van Hemert, "Understanding the Subprime Mortgage Crisis," SSRN paper, (December 5, 2008). Available at SSRN: http://ssrn.com/abstract=1020396.

[9.] Derrig, R. and L. Francis, "Distinguishing the Forest from the TREES: A Comparison of Tree-Based Data Mining Methods," *Variance* 2:2, 2008, pp. 184-208.

[10.] De Ville, B., *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*, SAS Publishing, 2006.

[11.] Ferris, Shauna, "How to Destabalize a Financial System: A Beginner's Guide," presented at the Biennial convention of the Institute of Actuaries of Australia, April 19-22, 2009.

[12.] Forray, S. "The Rat without a Tail," *Contingencies*, September/October 2009, pp. 32-38.

[13.] Gorton, G., "The Subprime Panic," National Bureau of Economic Research Working Paper, October 2008.

[14.] Helyar J., K. Burton, V. Silver, "Madoff Enabelers Winked at Suspected Front-Running," Bloomberg, Jan 27, 2009, http://www.bloomberg.com/apps/news?sid=au4Y7Cudw2Xo&pid=20601109.

[15.] Joint Risk Management Section (JRMS), *Risk Management: The Current Financial Crisis, Lessons Learned and Future Implications*, published as an e-book, December 2008, see http://www.casact.org/cms/pdf/Essays_on_Financial_Crisis.pdf.

[16.] Kedroski, Paul, "Berrnie vs. Benford's Law: Madoff Wasn't That Dumb," December 19, 2008, http://paul.kedrosky.com/archives/2008/12/19/bernie_vs_benfo.html.

[17.] LISC Research and Assessment, "Zip Codes Foreclosure Needs Methodology Appendix," October 2008, http://www.housingpolicy.org/assets/foreclosure-response/zipmethodology.pdf.

[18.] Markopolos, H., "Testimony of Harry Markopolos, CFA, CFE, Before the U.S. House of Representatives Committee on Financial Services," Wednesday, February 4, 2009, 9:30 AM, http://ncc.gmu.edu/events/09_MOT-markopols/markopolos_020409.pdf.

[19.] Morgenson, Gretchen, "Was There A Loan It Didn't Like?" *New York Times*, November 1, 2008, http://www.nytimes.com/2008/11/02/business/02gret.html.

[20.] Muolo, P. and M. Padilla, *Chain of Blame: How Wall Street Caused the Mortgage and Credit Crisis*, John Wiley and Sons, 2008.

[21.] Sanche, R., Kevin F. Lonergan, "Variable Reduction for Predictive Modeling with Clustering," CAS *Forum*, Winter 2006, pp. 89-100, http://www.casact.org/pubs/forum/06wforum/06w93.pdf.

[22.] Scheer, D., "SEC Never Did Competent Probe," Bloomberg News, September 2, 2009, http://www.bloomberg.com/apps/news?pid=20601087&sid=agBw9n2hZi5U.

[23.] Schoolman, P., "Credit Crisis Lessons for Modelers," in *Risk Management: The Current Financial Crisis, Lessons Learned and Future Implications*, 2008, http://www.soa.org/library/essays/rm-essay-2008-schoolman.pdf.

[24.] Triola, M., *Elementary Statistics*, 9th Ed., Addison-Wesley, 2003.

Biographies of the Authors

Louise Francis is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She is the 2009-11 CAS Vice President-Research & Development. She has won the Data Quality, Management and Technology call paper prize five times. She co-authored the paper "Actuarial I.Q. (Information

Quality)." She has also co-authored other papers on the subject of actuarial data management and data quality, as well as data mining.

Virginia Prevosto is a Vice President at Insurance Services Office, Inc. Ms. Prevosto has more than 30 years of actuarial and data management experience. Ms. Prevosto is a member of the American Academy of Actuaries, a Fellow of the Casualty Actuarial Society (CAS), and a member of the Insurance Data Management Association (IDMA) and the International Association for Information & Data Quality. Currently, she is chair of the CAS Committee on Management Data and Information, vice chairperson of the Joint Accreditation Committee, and a member of the Task Force on Basic Education Internet Modules. In the past, Ms. Prevosto also served as General Officer of the CAS Examination Committee; as chair of the CAS Candidate Liaison Committee; and as a member of the GIRO Data Quality Working Party and other professional committees of the CAS and AAA. Virginia has spoken at various venues on data management and data quality issues. Ms. Prevosto authored the paper "Study Note: ISO Statistical Plans," and co-authored the prize-winning paper "Dirty Data on Both Sides of the Pond" and the paper "Actuarial I.Q. (Information Quality)." She has also co-authored other papers on the subject of actuarial data management and data quality.

Acknowledgments

We wish to acknowledge Joseph Ng and Jason R. Smith, both who work for Insurance Services Office, Inc., for their invaluable assistance in data acquisition and preparation.