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

Louise Francis, FCAS, MAAA
Francis Analytics and Actuarial Data Mining, Inc
Virginia Prevosto, FCAS MAAA
ISO, Inc.

RPM Conference
March 16-17, 2010
Chicago, Il

Motivation

- Explore role of data in the financial crisis
- Illustrate that data was available
  - Much of analysis is exploratory
  - Some data mining will be illustrated
- Could have detected problems
  - Due diligence could have uncovered fraud
  - Provide warning of deterioration on mortgage quality

Two Case Studies of Use of Data to Detect Problems

- Madoff Ponzi Scheme
- Mortgage Crisis

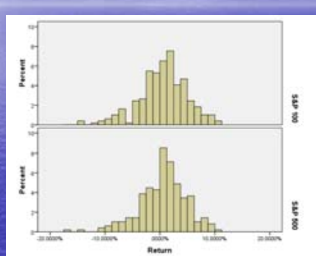
Madoff Ponzi Scheme

Could his fraud have been detected?
Should his data have been analyzed to verify that his returns were legitimate?

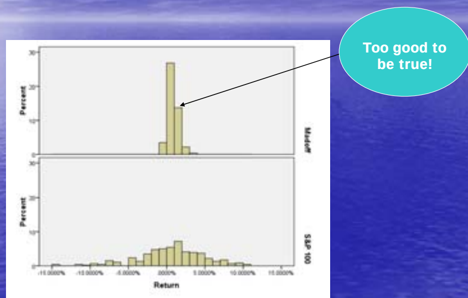
The data

- 1991 through 2008 returns on a Madoff feeder fund
- Downloaded from internet Jan, 2009
- This analysis motivated by Markopolis testimony to congress

Two similar assets: S&P 500 and S&P 100



Madoff vs S&P 100



Asset Descriptive Statistics

Statistics for Different Assets

Return

Name	Mean	Std. Deviation	Skewness	Kurtosis
Balanced	.43%	2.87%	-.89	1.54
Lng Bond	.67%	2.55%	.13	3.30
Madoff	*.83%*	*.70%*	*.77*	*.51*
S&P 100	.55%	4.39%	-.52	.84
S&P 500	.59%	4.31%	-.65	1.30
Total	.62%	3.39%	-.71	2.96

Percent of Time Negative Returns

Asset	Pct Negative Return
Balanced	39%
Lng Bond	37%
S&P 100	41%
S&P 500	38%
Madoff	*7%*

Min and Max

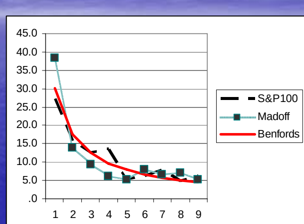
Asset	Median	Minimum	Maximum
Balanced	0.8%	-11.6%	5.7%
Long Bond	0.9%	-8.7%	11.4%
S&P 100	1.0%	-14.6%	10.8%
Madoff	0.7%	-0.6%	3.3%

Benford's Law

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%

Benford's law applied to Madoff data

- Usually applied to transactions
- Not a strong indicator of fraud applied to these returns



Madoff Case Study Conclusions

- Simple graphs and descriptive statistics could have detected the scheme
- Virtually all of them would have shown that the Madoff data deviates significantly from statistical patterns for similar assets

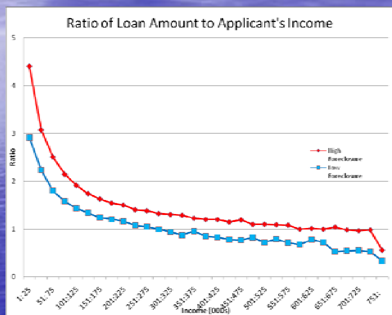
The Mortgage Crisis

Could simple descriptive statistics have predicted the meltdown?

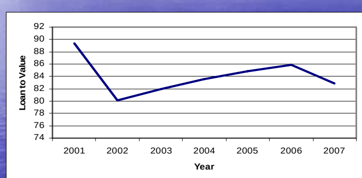
Some Descriptive Information from HMDA for Florida

		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

Ratio of Loan To Income

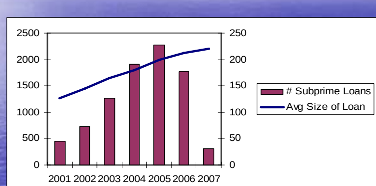


Time Series of Loan-to-Value



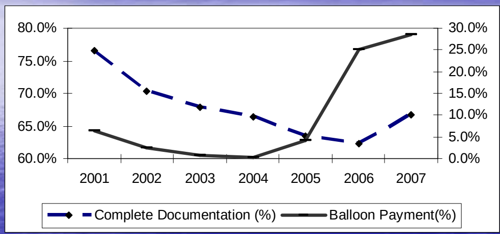
Data from Demyanyk and
Hemert, 2008

Subprime Loan Volume and Size



Data from Demyanyk and
Hemert, 2008

Balloon Payments and Completed Documentation



Data from Demyanyk and Hemert, 2008

Observations from HMDA

- HMDA indicates lower income applicants tend to have a higher loan to income ratio
- HMDA cross-state comparison indicates states with a foreclosure problem have consistently higher loan to income ratios compared to states not experiencing a foreclosure problem

Observations from Loan Portfolio Descriptive Statistics

- Subprime loans increased to unprecedented levels
- Loan to value increased
- Documentation decreased
- Balloon payments increased

Mortgage Fraud Analysis

Can data and models be used to detect mortgage fraud?

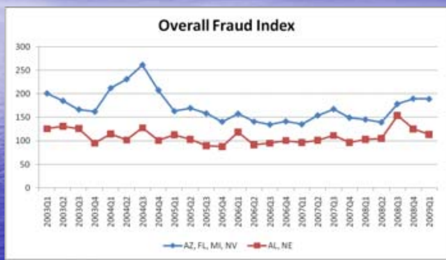
Interthinx Fraud Risk Index

- Uses detailed transaction data from loan applications processed by Interthinx's FraudGUARD System
- Uses relevant external data
  - Demographic, address data
  - Combination of methods

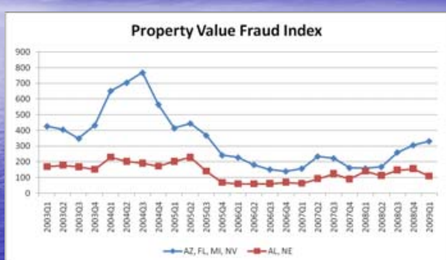
Subcomponents of Fraud Risk Index

- Property Value
  - Is appraisal value accurate?
- Identity
  - True identity of loan applicant? Is credit data accurate?
- Occupancy
  - Is applicant misrepresenting intent to occupy home?
- Income
  - Is income accurately stated?

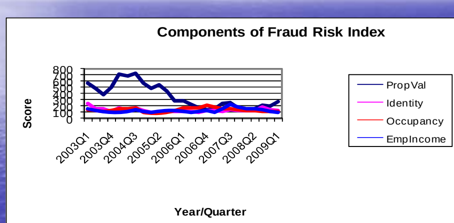
Overall Fraud Risk Index



Property Value Risk Index



Florida Subcomponents of Fraud Risk Index



Housing Data Trees

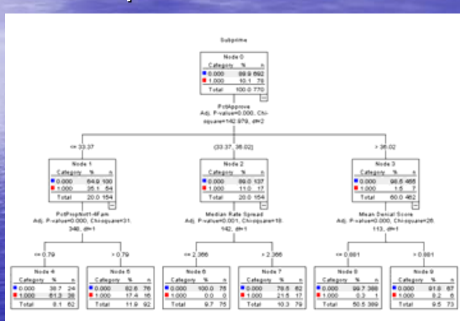
Could data mining have been used to predict subprime meltdown?

The Data

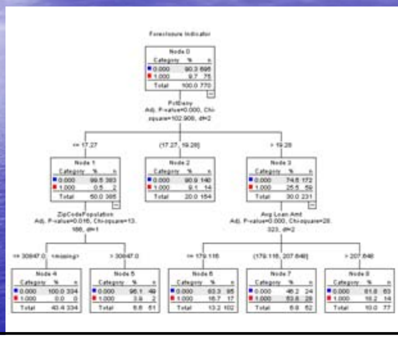
- HMDA Data
- LISC ZIP Foreclosure Needs Score
  - Subprime component
  - Foreclosure component
  - Disclosure component

  http://www.housingpolicy.org/foreclosure-response.html
- Zip Code Demographic Data

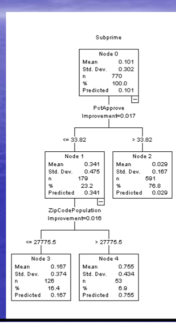
Subprime CHAID Tree



Foreclosure CHAID Tree



CART Subprime Tree



CART Foreclosure Variable Ranking

Independent Variable	Importance	Normalized Importance
Denial Percent	.027	100.0%
Mean Denial Score	.027	99.9%
PctApprove	.024	88.5%
ZipCodePopulation	.020	72.6%
PctPropNot1-4Fam	.019	69.5%
Median Rate Spread	.017	61.6%
PInCom	.016	60.5%
HouseholdsPerZipcode	.015	56.1%
Mean LTV Ratio	.014	52.7%

Results of Applying Clustering to HMDA Data

- K-means clustering applied to loan characteristics but not result data (i.e., approval)

Table III.5 – Means On Variables[1]

	Cluster		
	1	2	3
Avg Loan Amount	297.23	566.96	163.80
Average Income	165.71	356.66	87.26
Mean LTV[2] 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

Limitations of Data

- Origination Year vs Calendar Year

	Cumulative Default Rates @12/31/07								
	Development Age								
Year	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000
1999	0.013	0.076	0.131	0.179	0.202	0.223	0.231	0.236	0.239
2000	0.015	0.084	0.144	0.177	0.202	0.214	0.221	0.225	
2001	0.019	0.090	0.148	0.191	0.209	0.221	0.228		
2002	0.011	0.066	0.111	0.135	0.151	0.158			
2003	0.008	0.050	0.081	0.103	0.114				
2004	0.009	0.048	0.064	0.089					
2005	0.010	0.074	0.136						
2006	0.026	0.128							
2007	0.040								

Francis, L, "The Financial Crisis: An Actuary's View", in *Risk Management: The Current Financial Crisis, Lessons Learned and Future Implications*, 2008

Data Limitations

- As a result calendar year default rates are usually primarily attributable to earlier origination years
- It is likely that the 2007 default rates are largely driven by conditions in earlier years
- This affects interpretation of tree results

Observations

- Approval/Denial rate was an important variable for foreclosure and subprime problems
  - This may be a lagged effect. Low approval rates in 2007 reflect recognition of foreclosure problem originating in prior years when loose underwriting standards led to approval of risky and/or fraudulent loans
- Population and interest rate spread are additional important predictors of subprime problems
- Loan to income is an important predictor of foreclosures

Systemic Risk Data Collection Effort

www.ce-nif.org

- Questions?