STEVENS INSTITUTE OF TECHNOLOGY

GARP RESEARCH FELLOWSHIP

FALL 2018

Wavelet-based Time Series Cluster Analysis of Mortgage Risk

Author
Dhananjay Salgaocar
dsalgaoc@stevens.edu

Supervisor
Dr. Dragos Bozdog
dbozdog@stevens.edu

June 13, 2019

Author Cell Phone: 201-539-0713
Author Address: 121 1/2 Cottage Street #2,
Jersey City, NJ - 07306

Abstract

In this paper, we implement the discrete wavelet transform (DWT) to categorize mortgages based on the payment history. The wavelet transform is applied to the time series of payments to perform a multiresolution analysis. The resulting wavelet coefficients are used to cluster loans into three rating groups by using the various kMeans clustering methods. The first model proposed uses the wavelet coefficients corresponding to each scale to cluster the time series. The second model improves upon the first, by using an iterative procedure derived from the i-kMeans model that uses the final centers from clustering the higher scale coefficients to initialize the clustering for the next level. The third model makes use of the energy decomposition of wavelet coefficients to cluster the payment histories. It is seen that the k-Means algorithm fails to converge at lower scales when random centers are used. The i-kMeans algorithm addresses this problem by using cluster assignments from higher levels to initialize the centroids for the k-Means algorithm applied to lower level coefficients. The clusters are evaluated by observing the cluster assignments at the time of default. The clusters formed by the i-kMeans algorithm provide better separation than using individual-level coefficients while clustering based on energy decomposition manages to group all the defaults in a single cluster but has large numbers of false positives.

Contents

1 Introduction

Owning a home is an essential part of the "American Dream." Mortgages are debt instruments leveraged by home buyers to purchase property. If the borrower is not able to fulfill their obligations the lender can seize the property and liquidate it to clear the debt. As of the third quarter of 2018, the outstanding mortgage debt in the United States stood at 15.27 trillion dollars, of which one-to-four family residences hold 10.8 trillion. Approximately 64% of American homeowners have obtained a mortgage

Before entering a mortgage agreement lenders evaluate the creditworthiness of the borrowers to decide on the terms for the mortgage, or whether they should even lend the money or not. The evaluations are generally used to measure the ability of the borrower to pay back the loan. Lenders usually rely on the borrower's FICO score, debt-to-income ratio (DTI), proof of satisfactory income, employment, assets and what fraction of the home value they are willing to pay upfront.

Mortgages can vary based on the agreed-upon payment structure. Traditional mortgages are usually fixed-rate mortgages (FRM), where the monthly payments do not change in value throughout the mortgage. Most mortgages have a 15 or 30-year term. Another type of mortgage is the adjustable rate mortgage (ARM) where the interest rate charged changes over the mortgage term. The rate charged is usually tied to market interest rates. These mortgages may also have lower initial interest rates, making them more enticing to borrowers. Other types of mortgages include interest-only mortgages, balloon mortgages, and reverse mortgages.

Government Sponsored Entities(GSE) such as the Government National Mortgage Association (Ginnie Mae), the Federal Home Loan Mortgage Corporation (Freddie Mac) and Federal National Mortgage Association (Fannie Mae) foster mortgage lending by securitizing loans and selling these Mortgage Backed Securities (MBS) to investors. The creation of such instruments enables lenders to transfer the risk associated with mortgage lending to the buyers of these products, while also raising capital to finance more homeowners.

1.1 Objective

The probability of default is an estimate of the likelihood that a borrower may not be able to fulfill their financial obligations. Borrower and loan specific

variables are evaluated to determine the probability of default for a mortgage. While this information is available at the time of loan origination, it is likely to change over the duration of the mortgage. This information is usually updated irregularly and may not always reflect the borrower's creditworthiness.

The choice of a time series representation plays a significant role in obtaining an efficient and accurate model. Time series data poses many challenges for data mining due to the large size of data and high dimensionality. A large volume of data significantly increases the time taken to access data and perform analysis. Time series data mining may also suffer from the high dimensionality curse [1] that results in problems when using some distance measures. Time series data also exhibits different behavior when observed at different frequencies.

In this paper, it is hypothesized that similarities in the repayment patterns of mortgage loans may be used to develop a risk-rating model. In order to identify these similarities clustering models can be implemented. Since the data is observed monthly, time series clustering methods are suitable. The second part of the hypothesis is to test whether the Discrete Wavelet Transform can be used effectively to transform the time series data and solve issues related to the size, dimensionality and hierarchical nature of the data.

1.2 Research Approach

A literature review was conducted on current research in mortgage default. It is seen that most of the research is focused on borrower and loan specific variables. The use of the discrete wavelet transform for the statistical analysis of time series was also studied with a particular focus on clustering methods.

The Fannie Mae Single-Family Performance Data contains acquisition and performance data of a sample of mortgages on their books. A sample of loans was selected from the first quarter of 2011 to test the models.

Various wavelet-based models were used to cluster the time series. Since the discrete wavelet transform provides a time and frequency resolution of the input signal the coefficients are divided at corresponding frequencies. The energy preservation property of the wavelet transform was also leveraged. The models are applied over a rolling window of the payment histories.

The cluster assignments provided by the models were evaluated by aggre-

gating the cluster number just before the loan being foreclosed. The fraction of defaulted loans present in each cluster also offers insight into the general behavior of the cluster members.

1.3 Organization and Structure

Section 1 has presented the motivation behind this research and the proposed solution. The use of payment histories and the discrete wavelet transform for the transformation of time series has been suggested.

Section 2 will provide a review of relevant literature in mortgage research. The discrete wavelet transform will also be explained along with it's contributions to economic research. Following that, an introduction to time series clustering is described along with the k-Means algorithm. Examples of the use of wavelets in time series clustering are provided.

Section 3 describes the data set, processing of data, the general framework for the analysis and the models to be implemented. The models are divided according to the use of wavelet coefficients and their properties. First individual scale wavelet coefficients will be used to cluster mortgages, followed by an iterative procedure. The last type of method will make use of the energy decomposition of wavelet coefficients.

Section 4 is a collection of the results of the clustering models as well as a quantitative comparison of the outcomes of implemented methods.

Section 5 will summarize the paper and give suggestions on how the findings can be extended.

2 Literature Review

Since the 2008 Global Financial Crisis, much research has been done to analyze the creditworthiness of mortgages. Morrow [19] studies the effects of loan and borrower characteristics on the probability of default. The loans selected originated in 2006 and were studied up to 2012 with default being defined as being 180-days delinquent. The analysis includes logistic regression to infer the relationships between the variables and principal component analysis to determine which variables are most influential in mortgage default. They find that the interest rate charged and the loan amount is the most relevant variables and that all the other variables show the expected relationships with the probability of default.

While Morrow [19] studied variables that are recorded only at origination Ponomareva [22] introduces the idea that loans move through various creditworthiness profiles over their lifespan. Their study tests the accuracy of 4 multi-classification models: a basic baseline model, a deep neural network, a one dimensional LTSM recurrent neural network, and a one dimensional convolutional neural network. The models are evaluated on their ability to predict the delinquency status of loans 12 months into the future dependent on the payment history over the last year and origination variables. The study concludes that the other models show an improvement over the baseline model.

The notion of dynamically evaluating loans is extended by Sealand [25]. The study addresses default as a classification problem defining the dependent variable to indicate whether a loan will default over the next 12 months. Multiple models are trained on loan origination and performance data over a calendar year. Social, economic and financial variables are also involved in the subset of independent variables. The models are evaluated across multiple performance measures to check whether they can predict that a loan will default in future calendar years. Of the 15 models tested Boosting classifiers showed the best performance. When combining the predictions from multiple models, it was shown that ensemble decision trees were the most accurate models. It is interesting to note that when knn classifiers are optimized, they show a significant jump in accuracy. On average the models trained on the years from 2006 to 2016 proved suitable to predict default over the next two years after which the accuracy declined below the required threshold.

Another model that involves a dynamic approach to studying mortgage default available in Campbell and Cocco [5]. They propose studying the impact of macroeconomic variables such as labor income, house prices, inflation and interest rates to assert that mortgage default is triggered by negative home equity. They conclude that negative home equity plays a more significant role in the probability that loans with a higher Loan-to-Value ratio. They also show that interest rates affect both FRM and ARM loans while interest-only loans are least affected by negative home equity.

Foster and Van Order [8] proposed modeling a mortgage as a put option where mortgage default is treated as exercising the option. The borrower would trigger default when there is significant negative home equity but would have to overcome transaction costs. Elul et al. [6] extend upon this by defining borrowers with high credit utilization as illiquid. They then conclude that a combination of illiquidity and negative home equity lead to triggering a default, especially when the loans have higher Combined Loan-to-Value (CLTV) ratios.

2.1 Discrete Wavelet Transform

The Discrete Wavelet Transform is a mathematical tool that provides a time and frequency decomposition of an input signal or time series. It has found multiple uses in signal processing[23], image compression[15] and also biomedical data anslysis[26]. The DWT extracts time and frequency localized coefficients from an input signal or function that form its wavelet decomposition. An interesting property of the transform is that it preserves the energy of the time series. As a result of this the coefficients can be used to reconstruct the original signal. Another advantage of the DWT is that it can be used to detect anomalies in the signal. This is a result of its time localization property.

In order to understand the discrete wavelet transform it is first important to understand the concept of wavelets. A wavelet is essentially a small wave that grows and decays over a limited period of time. Mathematically, it is a real-valued function $\psi(.)$ that satisfies the following properties[21].

- The integral over $[-\infty, \infty]$ is zero:

$$\int_{-\infty}^{\infty} \psi(u)du = 0$$

- The integral of the square of the wavelet is 1:

$$\int_{-\infty}^{\infty} \psi(u)^2 du = 1$$

The simplest wavelet is the Haar wavelet that is defined as follows:

$$\psi(u) = \begin{cases} 1 & 0 < u \leqslant \frac{1}{2} \\ -1 & \frac{1}{2} < u \leqslant 1 \\ 0 & otherwise \end{cases}$$

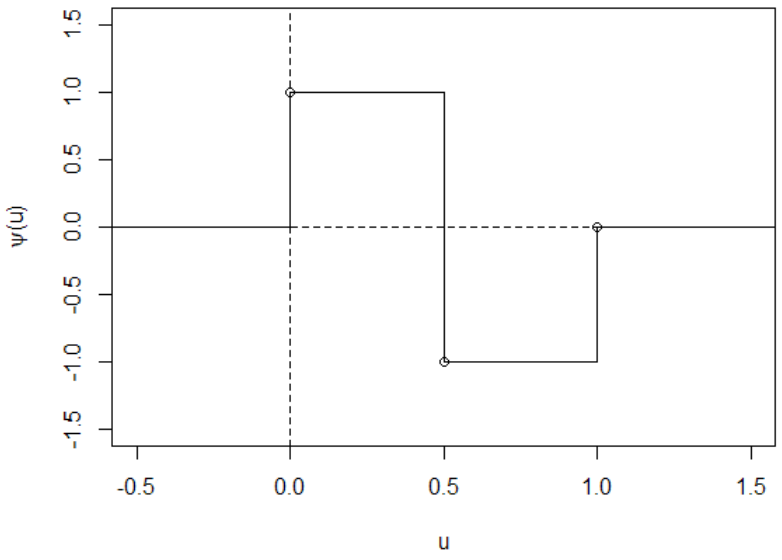

Figure 1: Haar Wavelet

Wavelets such as the Haar wavelet are used as the basis functions in wavelet transforms. While the use of wavelets allows us to capture localized changes in the time series a scaling function $\phi(u)$ is used to capture the broad approximation of the input function.

The Haar scaling function is given by:

$$\phi(u) = \begin{cases} 1, & 0 < u \leqslant 1 \\ 0, & otherwise \end{cases}$$

By scaling and translating the scaling and wavelet functions we form an orthonormal basis for the Hilbert space($\mathcal{H}$)

$$\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi(2^j t - k)$$

$$\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k)$$

where j is the scaling term while k is the translating term.

Any function $f \in \mathcal{H}$ can be represented as

$$f(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t)$$

where $j_0 \geqslant 0$. The coefficients $c_{j,k}$ and $d_{j,k}$ are called the scale and wavelet coefficients respectively.

Mallat [17] provides a fast algorithm to compute the multiresolution analysis of an input signal with length N where $N = 2^J$ and $J \in \mathbb{Z}$.

$$f_J(t) = c_0\phi_{0,0}(t) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t)$$

The set d_j represents the set of wavelet coefficients at the scale j

When using the Haar wavelet to obtain the wavelet transform, Mallat [18] showed that it is equivalent to calculating pairwise averages and differences. Assuming we have a time series S whose length N is of the order 2^j where j is some positive integer. The pairwise averages are $(S_{2n} + S_{2n-1})/2$ and differences are $(S_{2n} - S_{2n-1})/2$. For example:

Haar Wavelet Decomposition		
Level	Averages	Differences
S	$(53, 75, 25, 43, 52, 28, 35, 26)$	
1	$(64, 34, 40, 30.5)$	$(11, 9, -12, -4.5)$
2	$(49, 35.25)$	$(-15, -4.75)$
3	(42.125)	(-6.875)

Table 1: Haar Wavelet Decomposition

There has been a significant increase in the use of the discrete wavelet trans-

form in the analysis of financial time series. Gallegati [9] makes use of the frequency decomposition to show that stock returns lead economic activity at higher scales(lower frequencies). They indicate that wavelets have the potential to analyze relationships between processes that operate on different time scales.

In order to understand the causal relationship over time and scale between interest rates and stock prices in the Indian economy, Tiwari [27] made use of the multiresolution property of the wavelet transform. They were able to conclude that there was a causal and reverse causal relationship between the two variables at different time scales.

Hsieh et al. [11] makes use of the DWT to filter noise from the time series of stock prices. They integrate the wavelet transform and a recurrent neural network that uses fundamental and technical indicators to predict stock prices and develop a profitable trading strategy.

The use of wavelet analysis in mortgages is seen in González-Concepción et al. [10]. They make use of the frequency decomposition to study the relationship between the fraction of homeowners that have taken a mortgage and the gross domestic product in Spain. The wavelet analysis enabled them to show that the mortgage rate is positively correlated at lower scales(higher frequency) and leads the GDP while exhibiting a negative correlation at higher scales(lower frequency) with the GDP leading.

The Maximal Overlap Discrete Wavelet Transform(MODWT) is a modified version of the DWT. The MODWT can be directly derived from the DWT. It also allows for a multiresolution analysis. The MODWT also allows for signals of any length. The scaling and wavelet filters are defined as[21]:

$$\tilde{\phi}_{j,l}(t) = \phi_{j,l}(t)/2^{j/2}$$

$$\tilde{\psi}_{j,l}(t) = \psi_{j,l}(t)/2^{j/2}$$

where L is the length of the filter and $l = 1, 2, ..., L$. The scaling and wavelet coefficients are calculated in a similar fashion to the DWT by convoluting the filter and time series $X = \{X_t, t = 0, 1, 2....N-1\}$.

We first define $L_j = (2^j - 1)(L-1) + 1$. The matrices of scaling and wavelet coefficients can be calculated as follows.

$$\widetilde{V}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{\phi}_{j,l}(t) X_{t-l mod N}$$

$$\widetilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{\psi}_{j,l}(t) X_{t-l mod N}$$

The MODWT also preserves the energy of the signal.

2.2 Time Series Clustering

A time series is a sequence of values of some variable observed at various instances of time. Examples of time series are stock price data, daily temperature data and even medical data such as electrocardiograms. Generally, there are two kinds of methods applied to such data. Time series analysis is used to derive information from observed temporal data while time series forecasting is used to predict future values.

Clustering is a type of unsupervised statistical learning that involved collecting data that exhibit similar characteristics into groups called clusters. The most popular clustering methods are Hierarchical or Centroid-based. Hierarchical models rely on either a top-down or bottom-up approach to form a tree structure based on the distance between individual data points. In contrast to this, centroid-based models initialize a set of centroids that group the data effectively based on some distance measure.

One of the most popular centroid-based clustering algorithms is the k-Means algorithm[16]. The first step of the model is the selection of k which is the number of clusters. After this k centroids are initialized, either randomly or using known values. The distance of the data points from these centroids is determined using a suitable distance measure. The data points are assigned to the cluster represented by the centroid closest to them. After this, the centroids are recalculated by taking the mean of the data points in each cluster, and the process is repeated till a predefined number of iterations is reached, or the cluster assignments between iterations do not change. Since k-Means must converge to an optimum solution and that the convergence could be to a local rather than a global solution, the choice of initial clusters is highly influential.

Clustering time series data imposes many challenges due to the large size of time series data, high dimensionality, and difficulty in selecting a suitable similarity or dissimilarity measure. A variety of methods have been used to overcome these challenges. Examples of these methods are Dynamic Time Warping(DTW)[4], Correlation-based distances [15] and Euclidean distance [7].

The use of the Discrete Wavelet transform overcomes many of the difficulties in time series clustering. The wavelet coefficients can be used to reduce the size of the data. Santoso et al. [24] makes use of the transform by calculating a practical threshold level and only considering wavelet coefficients that are above the threshold. Lang et al. [14] uses the DWT as a noise filtering technique, also by thresholding the wavelet coefficients. By closely monitoring the coefficients at each scale Wang [28] describes a method to observe large jumps in simulated as well as real stock prices.

3 Data and Methodology

3.1 Fannie Mae Single-Family Loan Performance Data

The Fannie Mae Single-Family Loan Performance Data contains acquisition and performance data of housing loans acquired by Fannie Mae from the year 2000 onward. The loans are fully amortizing, single family and only fixed-rate. The data is available in the form of 2 text files that represent acquisition and performance loans for each quarter. For this project data from the first quarter of 2011 will be used.

The acquisition file contains information about the origination of the loan. Some interesting variables in this file are the borrower credit score which is the FICO score reported by various credit bureaus, Original Loan to Value ratio (OLTV) that represents how much money the borrower put down on the mortgage. An OLTV of 100% means that the borrower did not put any money down. The Debt-to-Income ratio (DTI) represents a ratio of the borrower's monthly credit payments to their income. The file also has a variable that contains the annual interest rate charged. The rate remains constant over the term of the mortgage as only fixed-rate mortgages are included.

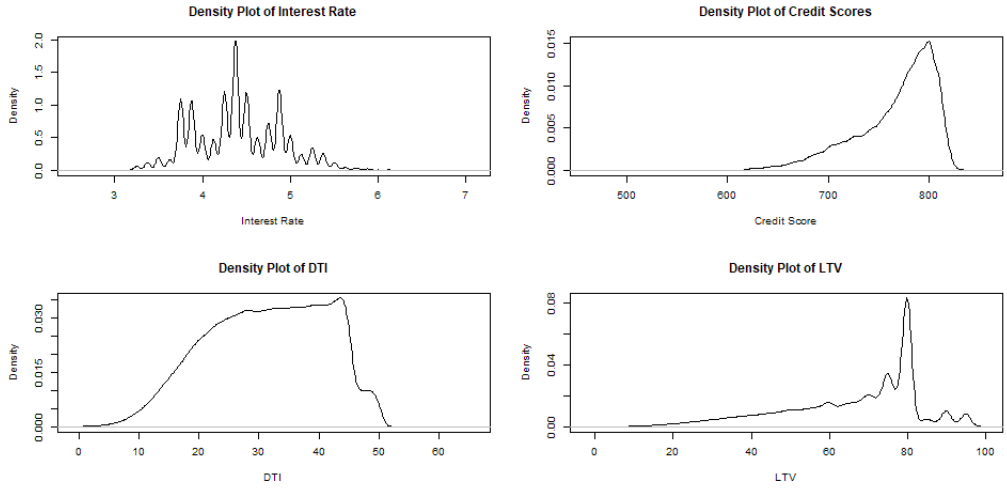

Figure 2: Distribution of variables in dataset

The performance file contains a monthly record of the loan's payment history and any other events that may have affected the status of the loan. In addition

to the monthly delinquency status, this file also contains the updated unpaid balance, loan age, and information about various credit events. Variables that describe the dates and costs associated with credit events such as prepayment, foreclosure, and modification are present in this file. Only the delinquency status sequences will be used in this study.

3.2 Data Processing

The dataset is available in the form of pipe delimited text files that were downloaded from the website. Each quarter was processed individually in R. Acquisition and performance data were combined to create a combined data set that reflects the latest available information such as the latest unpaid balance, the current status of the loan and if the loan is current, the delinquency status if the borrower has any late payments. The combined data set can be used to effectively filter loans that match specific criteria as the data contains a combination of acquisition and performance data. Each loan has a unique Loan ID that serves as the primary key across the tables.

A MySQL database containing loans that originated in the first quarter of 2000 until the third quarter of 2017 was designed. Each quarter has an acquisition, performance and combined table. The data was stored on a MySQL Server created on an Ubuntu Virtual Machine.

3.3 General Framework

The R programming environment was used to implement all the models described in this thesis. It provides a flexible framework as it supports various types of data and has a vast library of packages available. The 'RMySQL' package was used to interface with the MySQL database created to house the data. The package provides an easy pipeline to import data directly to an R data frame object. The 'wavelets' package was used to compute wavelet transforms. It also contains a library of wavelet filters that make it more versatile.

Performance data for loans that originated in the first quarter of 2011 is to be used. Only loans that are current or have defaulted will be considered for the models implemented. At the very minimum 84 months of performance data are available. As the discrete wavelet transform is limited to working with sequences

of length 2^n a rolling window of 32 months will be considered. While it is possible to evaluate time series of any length using the MODWT, in order to compare the methods the same window will be chosen. The clusters can be evaluated by checking what fraction of loans belonging to the cluster when default was imminent. The models to be evaluated have been described below.

It is important to note that loans that have defaulted do not have any records after foreclosure occurs. It is also seen that loans have an increasing delinquency status up to the point of foreclosure. In order to preserve these loans in the rolling window, the delinquency status variable is incremented by one every month.

3.4 Clustering DWT Coefficients

The Discrete Wavelet transform has many advantages for time series analysis. It is effective in data reduction, noise reduction, and multi-resolution analysis. The method described here takes advantage of the independence of coefficients belonging to different scale resolutions. The different scales can be used to observe hidden patterns that may exist in the evolution of mortgage delinquency. Use of this property has been demonstrated in [26] and [12]. Another advantage of the wavelet transform is that it does not assume the time series to be stationary.

Since the rolling window we are considering is 32 observations in length wavelet coefficients can be obtained at five different scales. The number of coefficients is halved at each level as the frequency is decreased. Each set of coefficients can be used for clustering to identify similarities in the payment patterns. In this model, K-Means clustering is carried out individually at each level of the wavelet decomposition. The Euclidean distance measure is used. The model is also extended using MODWT coefficients.

3.5 Modified I-kmeans algorithm

The I-kMeans algorithm [13] is an anytime algorithm for clustering time series data. The algorithm performs an initial random-centered clustering using the wavelet coefficients at the lower frequency decomposition of the signal. The centers from this initial clustering are projected at the finer level approximations and used as initial centers for clustering at these levels. The algorithm is repeated till the highest level is reached, or the cluster assignments remain constant between

the levels. The algorithm has been shown to reduce the run time of each level of clustering as the choice of initial centers leads to faster convergence of the k-means algorithm. It also results in a better quality of clustering.

In the original algorithm, the centers were projected between levels by duplicating them. The algorithm is modified by retaining the original cluster assignment and averaging the wavelet coefficients at the lower level. The averages are then used as initial centers at the new level, and clustering is performed. The algorithm terminates if cluster assignments do not change or if the highest level is reached.

input : The number of clusters k, Time Series data of length $l = 2^n$
output: Cluster assignments for individual time series in data

```
1  i ← n − 1
2  while i ⩾ 1 or Clusters do not change do
3      Coefficients ← DWT(data, i) ; // Returns DWT coefficients at scale
         i
4      if i == n − 1 then
5          Clusters ← k-Means(Coeff, rand) ;    // Performs k-Means with
             random centers
6      else
7          Centers ← Avg(DWT(Clusters, i));
8          Clusters ← k-Means(Coefficients, Centers) ;   // Performs k-Means
             with initial centers
9      end
10     i ← i − 1;
11 end
```

Algorithm 1: Modified i-kMeans

3.6 Energy Based Clustering

Since feature extraction is a necessity when performing time series analysis the model proposed in [3] makes use of the energy decomposition of the discrete wavelet transform. Parseval [20] states that the integral of the square of the signal

is equal to the squared sum of its transform

$$\int |f(t)|^2 dt = \sum_{k=0}^{2_0^j-1} c_{j_0,k}^2 + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}^2$$

From this, a vector of energy contributions is constructed by studying the absolute and relative contributions of scales to the global energy. The absolute contribution of a scale j is given by

$$abs_j = \sum_{k=0}^{2^j-1} d_{j,k}^2$$

while the relative contribution is calculated by

$$rel_j = \frac{abs_j}{\sum_{j=0}^{J-1} abs_j}$$

It is seen that the approximation coefficients $c_{0,0}$ are left out of the contributions. The distribution of relative contributions can be used for distance based clustering as the sum of coefficients sums to 1.

4 Results

In this section, the results of applying the clustering methods described in the last chapter are discussed. The chapter is divided based on the type of feature extraction used for clustering. The first subsection discusses the use of separate frequency resolutions of the DWT and MODWT coefficients. The next subsection evaluates the i-kMeans algorithm that also made use of the DWT and MODWT followed by the results of clustering using the energy decomposition of the DWT coefficients. Lastly, the results are summarized.

The data set used for clustering was a sample of The Fannie Mae Single-Family Loan Performance Data. Twenty samples of five hundred loans each were used to implement the models described in the previous chapter. Each sample contained four hundred current loans and one hundred loans that default.

We define "at-risk" loans as loans that are one month away from default. As described in the previous chapter we have used a rolling window that clusters the loans based on thirty two observations of the time series. The results are presented by creating a distribution of cluster assignments of at-risk loans. In addition, we define the precision of the clusters as the ratio of defaulted loans assigned to the cluster to the number of loans in the cluster.

4.1 Clustering DWT Coefficients

The models that clustered loans based on the individual scale coefficients are evaluated in this section. At the fourth level (lowest frequency) we obtain 2 wavelet coefficients for each loan. The frequency associated with these coefficients is 16 observations of the time series. It is seen in Figure 3 that there is very little separation between clusters one and two. Cluster two also has a high variance in precision.

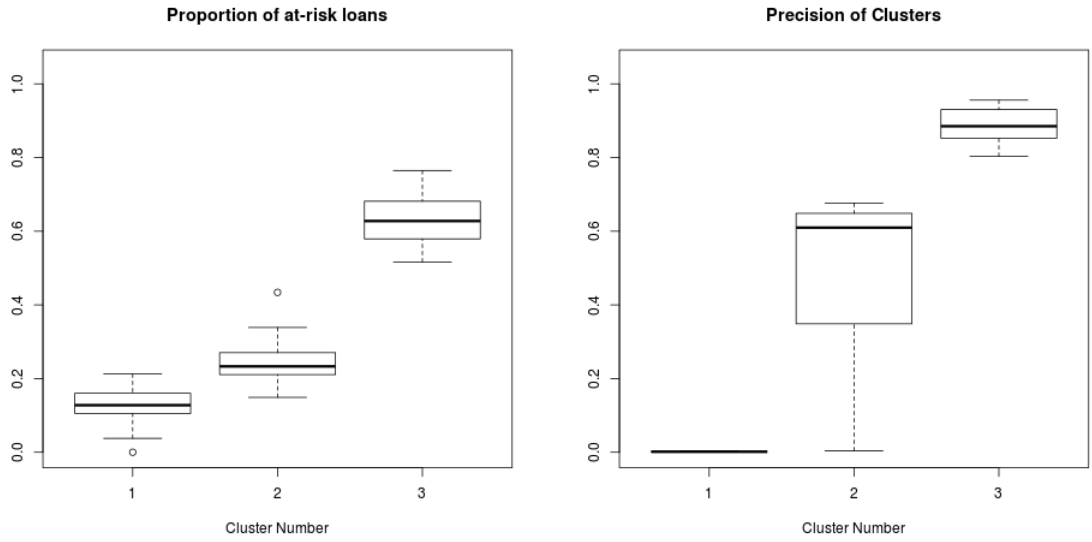

Figure 3: Distribution of At-risk Loans and Cluster Precision: DWT Level 4

When level 3 coefficients were used to cluster the loans it is observed that cluster one has a lower proportion of at-risk loans. Cluster three captures a higher proportion of such loans when compared to the level four coefficients and also provides better precision.

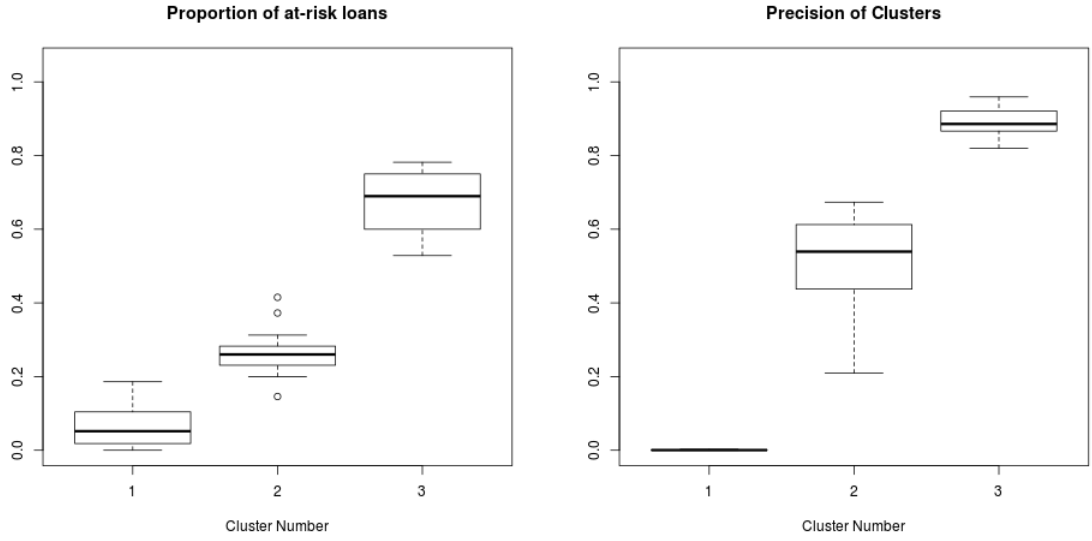

Figure 4: Distribution of At-risk Loans and Cluster Precision: DWT Level 3

Figure 5 shows the use of the level 2 coefficients for clustering. Almost all the

loans in cluster three default. There is not much difference between the capture of at-risk loans in clusters one and two. It is also worth noting that the k-Means model did not always converge when using these coefficients. Since each coefficient at this level approximates four observations of the time series we have eight such coefficients.

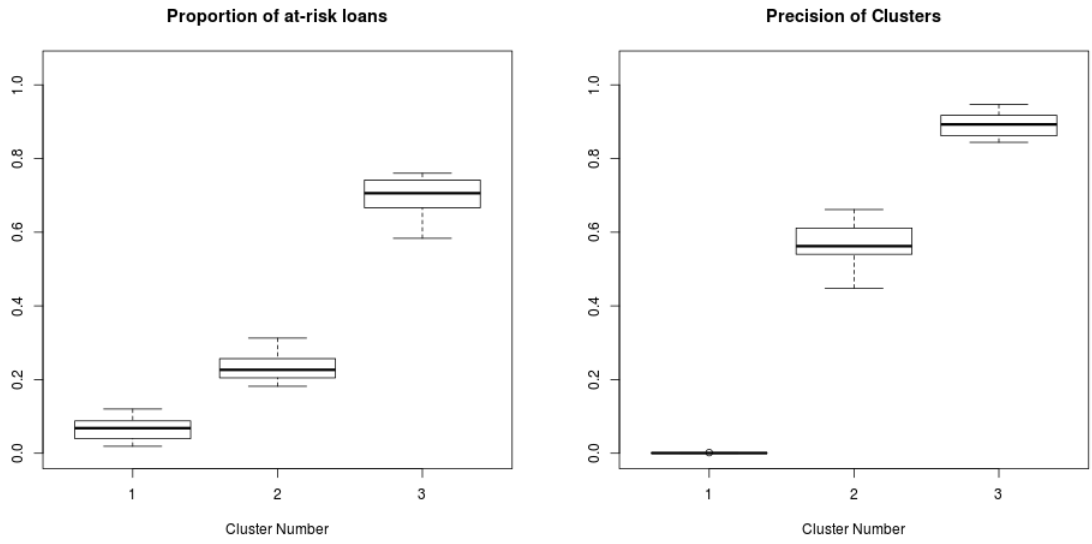

Figure 5: Distribution of At-risk Loans and Cluster Precision: DWT Level 2

At the first level of wavelet coefficients, each coefficient encompasses 2 consecutive observations of the time series. The results are similar to the level two coefficients. The k-Means did not always converge at this level either.

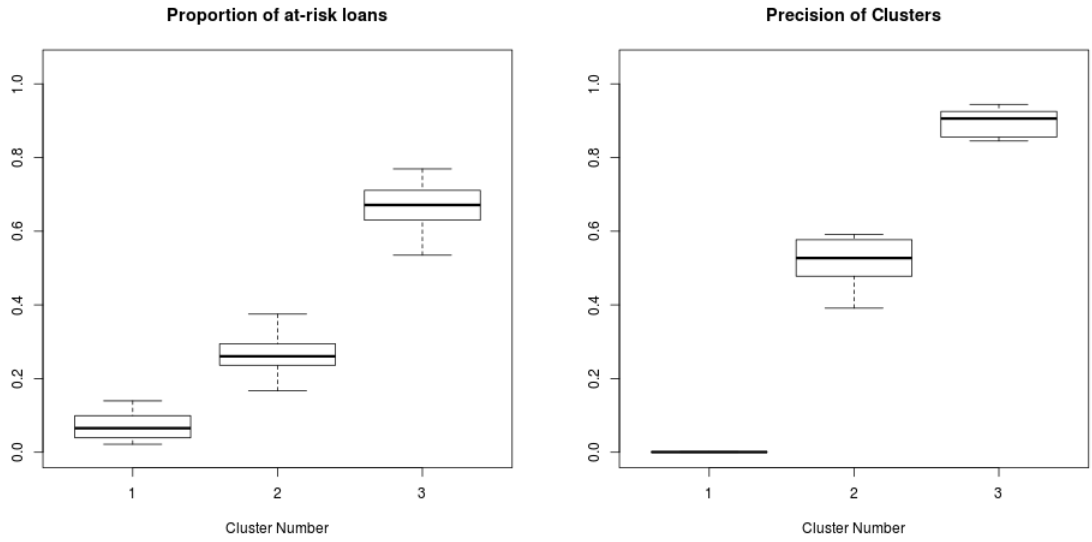

Figure 6: Distribution of At-risk Loans and Cluster Precision: DWT Level 1

When using MODWT coefficients for clustering it is observed that there is very little separation between the clusters when evaluated for capturing at-risk loans. The distribution of precision shows a similar results to the use of DWT coefficients.

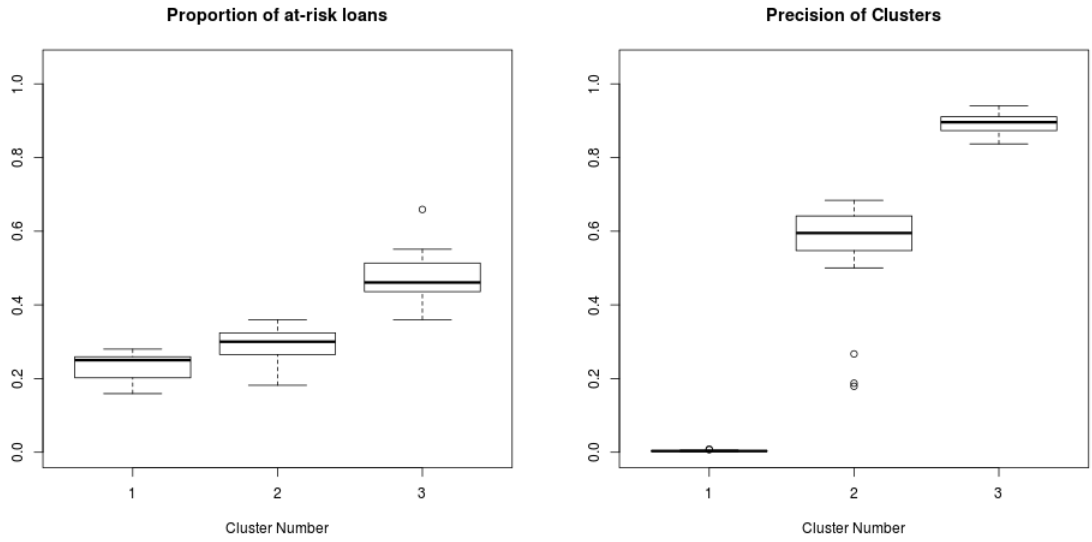

Figure 7: Distribution of At-risk Loans and Cluster Precision: MODWT Level 4

The use of level 3 coefficients does show a slight improvement in capturing

loans that are on the verge of default. However the MOWDT coefficients do not have any benefit over the DWT coefficients.

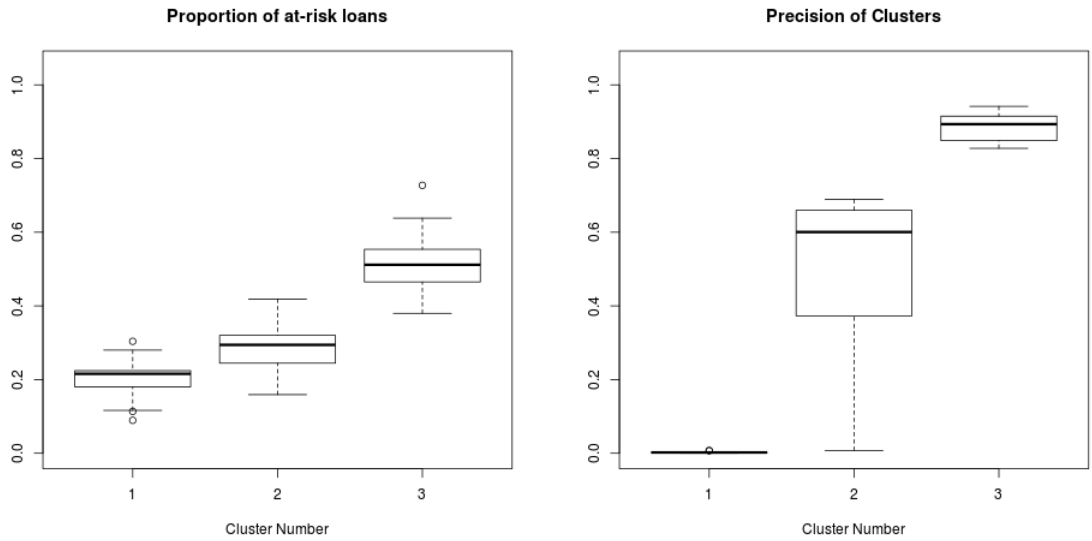

Figure 8: Distribution of At-risk Loans and Cluster Precision: MODWT Level 3

When using the MODWT coefficients at levels 1 and 2 the k-Means algorithm did not converge for a majority of the samples.

4.2 Modified I-kMeans algorithm

In this section the results of clustering using the i-kMeans algorithm are described. The algorithm was implemented using DWT as well as MODWT coefficients.

The use of DWT coefficients for the i-kMeans algorithm results in clusters similar to those of clustering with the level 1 coefficients. Cluster 3 captures approximately 60% of the loans that are near default. It is observed that when the centers or k-Means are randomly initialized, the convergence of clustering with level 1 coefficients requires more iterations than compared to the level 1 clustering in the i-kMeans algorithm.

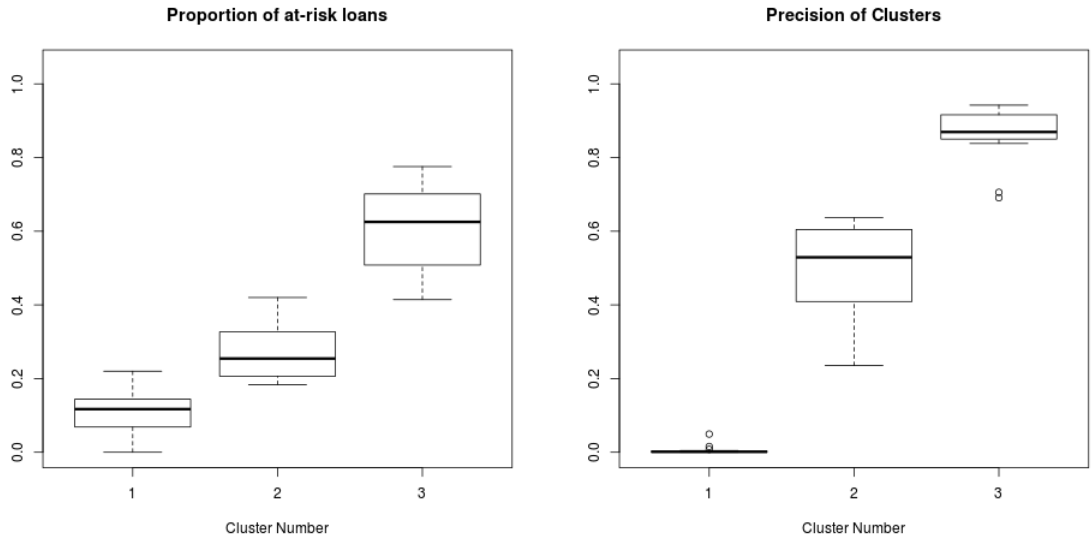

Figure 9: Distribution of At-risk Loans and Cluster Precision: DWT i-Kmeans

When using MODWT coefficients in the i-kMeans algorithm, the clusters capture the at-risk loans more evenly. The i-kMeans algorithm also results in convergence of clustering for level 1 & 2 coefficients of the MODWT.

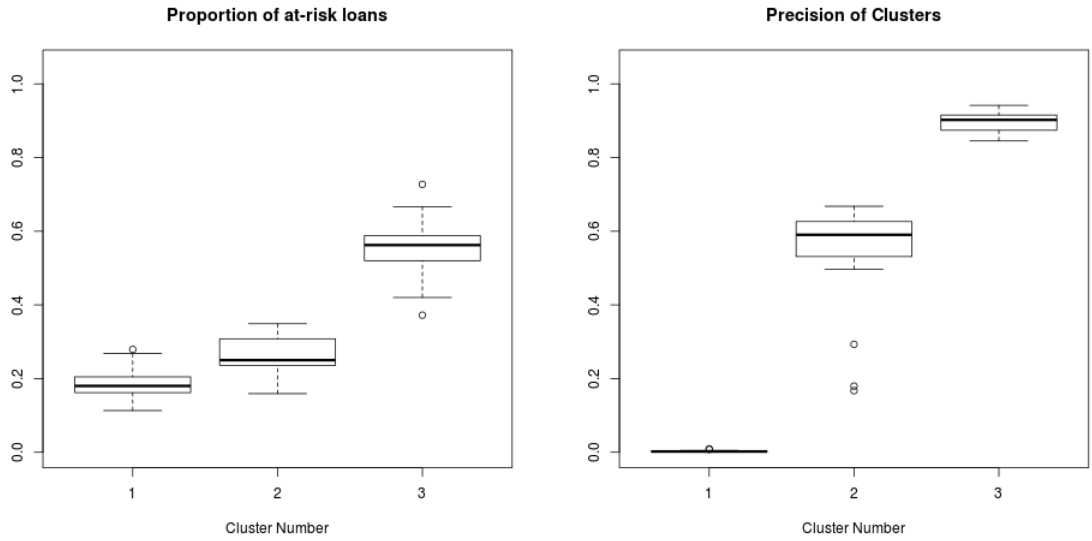

Figure 10: Distribution of At-risk Loans and Cluster Precision: MODWT i-Kmeans

4.3 Energy Based Clustering

In this section results of clustering loans based on their energy decomposition are discussed. The clustering is first tried using 3 clusters and also with $k = 2$.

When the loans are grouped in 3 clusters it is seen that the third cluster captures almost all of the loans that are on the verge of defaulting. Clusters 1 does not capture any. The ratio of defaults to loans clustered is also very high for cluster 3 while being close to 0 for the others.

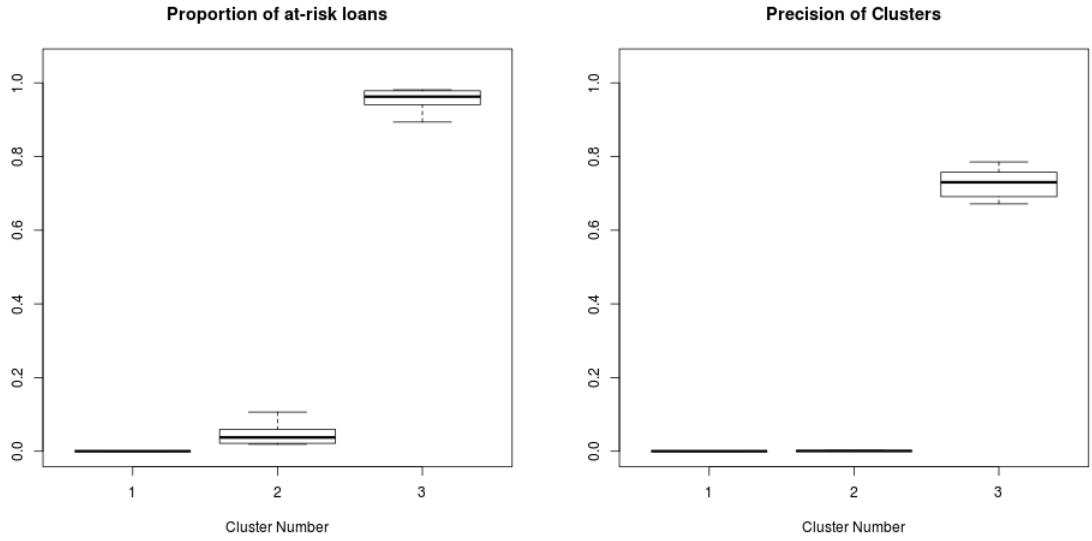

Figure 11: Distribution of At-risk Loans and Cluster Precision: Energy Distribution with $k = 3$

Based off the results of clustering with $k = 3$, when the number of clusters is set to 2 we get one cluster that captures all the at-risk loans. However we get two clusters that are of almost of equal size which results in very low precision.

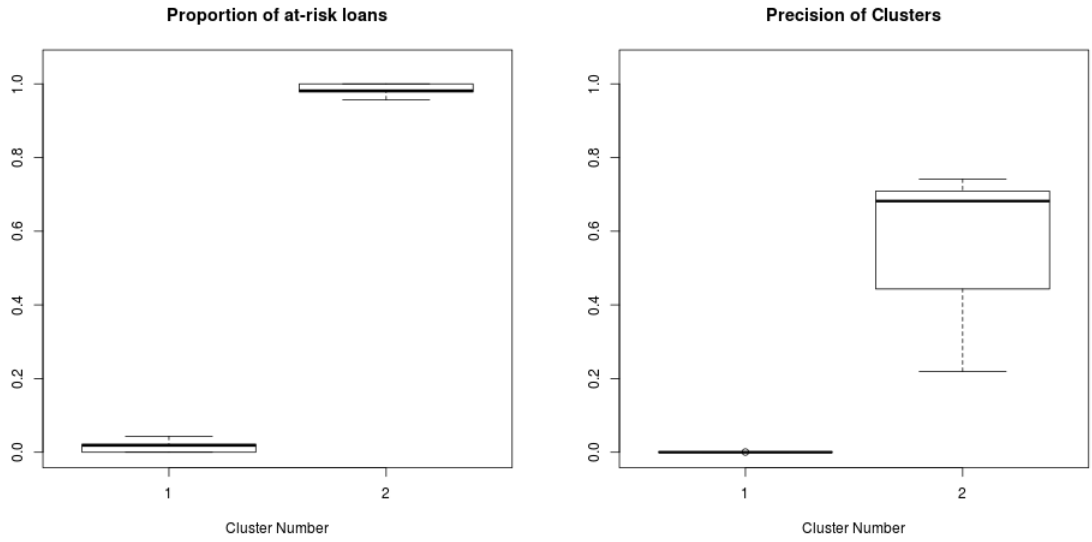

Figure 12: Distribution of At-risk Loans and Cluster Precision: Energy Distribution with $k = 2$

4.4 Results Summary

In this section the results from each model are summarized. The clustering has been evaluated by observing the proportion of loans that are grouped together just before they are foreclosed on and also by checking what proportion of loans in each cluster have actually defaulted. Table 2 shows the aggregation of cluster assignments prior to default as a proportion of the total number of loans that default.

It was seen in the previous section that the energy based clustering method captured most of the at-risk loans in a single cluster. When the number of clusters is reduced from 3 to 2, all the loans with imminent default are in a single cluster. Table 2 confirms these results. The clusters with the most at-risk loans contained 92% and 100% of the loans that were about to go into default.

	Mean			Standard Deviation		
Model	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
DWT4	0.1234	0.2452	0.6313	0.0516	0.0666	0.0721
DWT3	0.0659	0.2667	0.6674	0.0634	0.0630	0.0877
DWT2	0.0670	0.2363	0.6967	0.0352	0.0400	0.0555
DWT1	0.0693	0.2657	0.6650	0.0376	0.0586	0.0688
MODWT4	0.2340	0.2944	0.4717	0.0350	0.0441	0.0691
MODWT3	0.2016	0.2856	0.5128	0.0555	0.0665	0.0862
DWT i-kMeans	0.1121	0.2765	0.6114	0.0561	0.0761	0.1141
MODWT i-kMeans	0.1863	0.2603	0.5534	0.0505	0.0519	0.0859
Energy k=3	0.0000	0.0422	0.9578	0.0000	0.0236	0.0236
Energy k=2	0.0145	0.9855	NA	0.0154	0.0154	NA

Table 2: At-risk Loan Distribution Statistics

Since the models are applied over a rolling window the precision of models is observed over time. Table 3 summarizes the mean and standard deviation of each cluster's precision.

	Mean			Standard Deviation		
Model	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
DWT4	0.0015	0.4922	0.8849	0.0009	0.2225	0.0498
DWT3	0.0007	0.5135	0.8921	0.0007	0.1251	0.0369
DWT2	0.0006	0.5682	0.8913	0.0005	0.0611	0.0347
DWT1	0.0007	0.5169	0.8973	0.0004	0.0692	0.0357
MODWT4	0.0038	0.5475	0.8927	0.0017	0.1535	0.0290
MODWT3	0.0024	0.4868	0.8867	0.0020	0.2432	0.0400
DWT i-kMeans	0.0048	0.4946	0.8668	0.0111	0.1257	0.0662
MODWT i-kMeans	0.0028	0.5374	0.8966	0.0023	0.1536	0.0286
Energy k=3	0.0000	0.0008	0.7272	0.0000	0.0005	0.0351
Energy k=2	0.0001	0.5750	NA	0.0001	0.2053	NA

Table 3: Precision Statistics

When comparing the methods according to at-risk loan captures it is seen that the energy based clustering methods perform better than those based on the wavelet coefficients. However, the methods based on wavelet coefficients have a better precision.

5 Conclusion

5.1 Summary

In this paper, the idea that loans can be categorized into various risk levels based on their payment histories is tested. As it is possible to find similarities in payment patterns using clustering, k-Means clustering is implemented. The algorithms are tested on twenty samples of 500 loans that originated in the first quarter of 2011 from the Fannie Mae Single-Family Loan Performance Data. The loans were clustered into 3 clusters that could represent risk levels. This is demonstrated by using the results of the analysis to show that the clusters generally contained different ratios of loans that go into default. It is also seen that it can form clusters that contain loans that are about to default.

The use of the Discrete Wavelet Transform for feature extraction is also implemented. Three models that make use of individual-level wavelet coefficients, the entire transform, and the energy distribution between the levels of coefficients, are used to cluster the time series of loans. It is concluded that the modified i-kMeans algorithm ensures convergence of the clustering algorithm at lower levels of the wavelet transform coefficients. The energy decomposition performs well when used to group together loans that are on the verge of going into default.

An interesting result is that the clusters formed by the level 3 coefficients display the highest variance between the number of loans captured by the clusters when default was imminent. The ratio of defaulted loans in the third cluster was also on the higher side. This could indicate a relationship between default and the evaluation of loans at a frequency of 4 observations.

While the Maximal Overlap Discrete Wavelet Transform has proven to be useful in the analysis of time series[2], it does not have benefits for this application. All the models that incorporated the MODWT showed lower variance between the capture of loans about to default as well as a lower ratio of defaulted loans in the higher risk clusters.

5.2 Contributions

The main goal of this paper was to develop a framework that categorized loans based on how likely they were to default. Several wavelet-based cluster-

ing methods were evaluated on a proposed performance measure. While wavelet methods have been used in finance, their applications on housing loans have not been tested.

Two unique methods of using wavelet coefficients for statistical analysis have been implemented. The first is an iterative procedure that makes use of the coarser approximations to make clustering algorithms faster and more reliable. The other takes advantage of the energy retention property of wavelet transforms to generate features that can be used in all types of statistical learning.

Although the use of the MODWT has been shown to have an advantage over the DWT as it can be used on signals of any length, it has not proven to have any advantage for the categorization of housing loans.

5.3 Future Work

As estimating the probability of default may be the eventual goal of research in this field the applications of the DWT to time series classification could also prove useful. The energy distribution could be used to create a set of features for this. Since the output of classification algorithms can be interpreted as a probability, it can be used to evaluate the risk factor associated with default.

The ratio of loans that have defaulted in a cluster to the size of a cluster can be interpreted as a probability of default given that a loan is assigned that cluster. Such a system is similar to Hidden Markov Models (HMMs) that make use of state probabilities and emission probabilities. The cluster assignments can be used as the hidden states, and the other probabilities can be calculated.

These models can also be extended by making use of the origination variables. The initialization of cluster centers could incorporate these variables. Although updated borrower and loan characteristics are not available in this particular data set, if provided they could be important features for the risk-rating system.

References

[1] Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In *International Conference on Foundations of Data Organization and Algorithms*, pages 69–84. Springer.

[2] Alarcon-Aquino, V. and Barria, J. (2009). Change detection in time series using the maximal overlap discrete wavelet transform. *Latin American applied research*, 39(2):145–152.

[3] Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(01):1350003.

[4] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

[5] Campbell, J. Y. and Cocco, J. F. (2015). A model of mortgage default. *The Journal of Finance*, 70(4):1495–1554.

[6] Elul, R., Souleles, N. S., Chomsisengphet, S., Glennon, D., and Hunt, R. (2010). What" triggers" mortgage default? *American Economic Review*, 100(2):490–94.

[7] Faloutsos, C., Ranganathan, M., and Manolopoulos, Y. (1994). *Fast subsequence matching in time-series databases*, volume 23. ACM.

[8] Foster, C. and Van Order, R. (1984). An option-based model of mortgage default. *Housing Fin. Rev.*, 3:351.

[9] Gallegati, M. (2008). Wavelet analysis of stock returns and aggregate economic activity. *Computational Statistics & Data Analysis*, 52(6):3061–3074.

[10] González-Concepción, C., Gil-Fariña, M. C., and Pestano-Gabino, C. (2012). Using wavelets to understand the relationship between mortgages and gross domestic product in spain. *Journal of Applied Mathematics*, 2012.

[11] Hsieh, T.-J., Hsiao, H.-F., and Yeh, W.-C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied soft computing*, 11(2):2510–2525.

[12] Huang, P., Feldmann, A., and Willinger, W. (2001). A non-instrusive, wavelet-based approach to detecting network performance problems. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 213–227. ACM.

[13] Keogh, M. V. J. L. E. and Gunopulos, D. (2003). A wavelet-based anytime algorithm for k-means clustering of time series.

[14] Lang, M., Guo, H., Odegard, J. E., Burrus, C. S., and Wells, R. O. (1996). Noise reduction using an undecimated discrete wavelet transform. *IEEE Signal Processing Letters*, 3(1):10–12.

[15] Lewis, A. S. and Knowles, G. (1992). Image compression using the 2-d wavelet transform. *IEEE Transactions on image Processing*, 1(2):244–250.

[16] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[17] Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

[18] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):674–693.

[19] Morrow, J. (2015). Financing the american dream: Using logistic regression and principal component analysis to identify the probability of default in mortgage lending.

[20] Parseval, M.-A. (1806). Mémoire sur les séries et sur l'intégration complète d'une équation aux différences partielles linéaires du second ordre, à coefficients constants. *Mém. prés. par divers savants, Acad. des Sciences, Paris,(1)*, 1:638–648.

[21] Percival, D. B. and Walden, A. T. (2006). *Wavelet methods for time series analysis*, volume 4. Cambridge university press.

[22] Ponomareva, K. (2018). Predicting mortgage loan delinquency status.

[23] Rioul, O. and Vetterli, M. (1991). Wavelets and signal processing. *IEEE signal processing magazine*, 8(ARTICLE):14–38.

[24] Santoso, S., Powers, E. J., and Grady, W. (1997). Power quality disturbance data compression using wavelet transform methods. *IEEE Transactions on Power Delivery*, 12(3):1250–1257.

[25] Sealand, J. C. (2018). *Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models*. PhD thesis, Slippery Rock University.

[26] Subasi, A. (2005). Epileptic seizure detection using dynamic wavelet network. *Expert Systems with Applications*, 29(2):343–355.

[27] Tiwari, A. K. (2012). Decomposing time-frequency relationship between interest rates and share prices in india through wavelets.

[28] Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397.