Clustering of Time Series using Wavelet Transformations as a Feature Extraction Mechanism

Kathy Norman, Ssurey Moon, Felix Huang, Josué Kuri
UCSC Extension. 30164:(003) Machine Learning and Data Mining

May 30, 2015

1 Introduction

A time series is a sequence of data points indexed by time at regular intervals. This model is used to represent a wide range of metrics such as the daily closing price of stocks, temperature, precipitation, population, etc. In the context of machine learning and data mining, clustering of a large set of time series is an exploratory technique aimed at identifying and understanding underlying patterns.

Considering every point of a time series as a dimension results in a high dimensional space which clustering algorithms cannot handle easily. These algorithms depend on a distance measure as a basis to maximize cohesion and separation. In a high dimensional space, the contrast between the nearest and the farthest neighbor becomes smaller making it difficult for clustering algorithms to find meaninful groups [1].

Data dimensionality reduction is an approach to map a high dimensional space into a lower dimensional space such that the main characteristics of the data points in the original space are preserved and clustering on the lower dimensionality space results in meaningful groups. The two types of dimensionality reduction are feature selection and feature extraction. The former consists in selecting a subset of features from the original features. Feature extraction, on the other hand, generates a new set of features through a mapping function.

Feature extraction techniques commonly used include Singular Value Decomposition (SVD), Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). Of these techniques, SDV is the most effective at reconstructing time series with minimal error. However, its time complexity $O(mn^2)$, where m is the number of time series and n is the lenght of each time series, makes this a computally-intensive approach [2]. A Fast Fourier Transform (FFT) algorithm can compute DFT coefficients in $O(mn \log n)$ and DWT, using a spacial type of wavelet called *Haar wavelet* can achieve $O(mn)$ [2].

In this project we use a feature extraction approach based on DWT using the Haar wavelet as the basis for the transformation. We create a generic framework for clustering of time series using this approach and apply the framework to three types of time series: daily closing stock prices, daily values of exchange rates, and earthquake activity over time for various geographic regions. We use a silhouette coefficient as the criterion to evaluate the quality of the clusterings generated by the framework.

2 Wavelet transformation

Wavelet transformation is a time-frequency domain transformation technique for hierarchical decomposition of signals [3, 4]. The decomposition creates an approximation of the original signal that preseves the trend of the signal, as well as additional data sets that provide increasing levels of detail to reconstruct the original signal. This original signal can be reconstructed without loss of information by applying an inverse wavelet transform to the combination of the approximation signal and all the detail data sets.

Early work on wavelets originated with Morlet in the 1980s as a new tool for seismic signal analysis [5]. Further work by Morlet, Grossman, Meyer, Mallat and Daubechies [6, 7] broght the concept to the mainstream mathematics community with applications in signal processing, statistics and other areas. There is at present a vast body of literature about the foundations and applications of wavelets. The interested reader is referred to [8] and similar works for a comprehensive presentation of the field.

3 Wavelet-based feature extraction

Consider a time series $\overrightarrow{X} \in \mathbb{R}^n$ as an ordered sequence of $n \in 2^J, J \in \mathbb{N}$ numbers. After decomposing $\overrightarrow{X}$ at a resolution $r \in \{2, 4, 6, 8, 10, ...\}$, the coefficients associated to the r level can be represented as a sequence $\{A_r, D_r, D_{r-1}, ..., D_2, D_1\}$. The first element A_r is the approximation coefficients array, and the subsequent elements $D_r, D_{r-1}, ..., D_2, D_1$ are the details coefficients arrays. In this project we use the vector $\widehat{X} = \{A_r, D_r\}$ at specific levels $r \in \{2, 4, 6, 8, 10\}$ as feature vectors for the clustering of a set of time series. The cardinality $|\widehat{X}|$ decreases as r increases. It results, on one hand, in a reconstruction with less fidelity than the original signal $\overrightarrow{X}$ and, on the other hand, on clusterings of potentially better quality because of the smaller dimensionality of $\widehat{X}$.

4 Experimental evaluation

The purpose of the experimental evaluation of the generic framework for clustering of time series is to assess the relative efficiency of the framework on a set of contrasting application domains ranging from financial markets to geological phenomena. Variations in the cardinality $|\widehat{X}|$ of feature vectors, the choosen clustering algorithm, as well as the number of clusters, are considered to assess what combination of algorithm and input parameters is the most appropriate for specific situations. The following clustering algorithms, implemented in the Scikit Learn python machine learning library [11], are used in the evaluation:

- Mini-Batch K-Means
- Affinity Propagation
- Spectral Clustering
- Ward
- Agglomerative Clustering
- Birch
- DBSCAN

4.1 Evaluation criteria

There is no explicit way to evaluate the quality of time series clustering methods since clustering is an unsupervised learning approach. In supervised learning data sets all data points are labelled. These labels (or classes) are used to calculate the SSE (Sum of Square Errors) or other measures of the quality of the implementation. In contrast, the evaluation of clustering implementation on unsupervised learning data set need another approach to internally calculate the quality of derived clusters. The most popular method is to get cohesion and separation. Cohesion means how close each node in a cluster $\sum_i \sum_j (x_{i,j} - m_j)^2$ is, where i is number of clusters, $x_{i,j}$ is a j_{th} node in the cluster i, and m_i is the center of the cluster i. On the other hand, separation means how far each cluster is from others, $\sum_k \sum_i (m_k - m_i)^2$. The higher those values are, the better nodes are clustered.

The silhouette coefficient combines the concepts of cohesion and separation:

$$s = \{ \, 1 - a/b \;\text{ for } a \leq b, \; b/a - 1 \;\text{ otherwise} \tag{1}$$

where a is the average distance x_i, a random node in the cluster i, and other nodes in the cluster, b is the minimum value the average distances of x_i and nodes in another cluster k. The closer to one the value is, the better nodes are clustered.

In this project we use Silhouette coefficient to measure the performance of our implementation. There are a few parameters to consider, such as clustering algorithm, level of complexity, in other words, feature extraction level, and a forced number of clusters. Silhouette score plots point which combination of parameters performs better on different type of time series data sets.

4.2 Data description

4.2.1 Stock closing prices

This data set consists of the time series of daily closing prices for one hundred stocks from diverse industries (Internet, telecommunications equipment and service providers, entertainment, media, airlines, etc.) between April 20th 2011 and May 16th 2015. The time series represent 1024 trading days, and are normalized using as basis the closing price of the first day of the interval (April 20th 2011). The complete list of stocks used in thi project is given in Appendix A.

4.2.2 Historic exchange rates

This data set consists of the exchange rate time series for thirty pairs of currencies, e.g., JPY/USD, over 1024 days. The time series are normalized using as basis the exchange rate on the first day of the interval. These time series are retrieved from the Federal Reserve Economic Data (FRED) service of the Federal Reserve Bank of St. Louis [10].

4.2.3 Historic earthquake data

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4.3 Performance evaluation

4.4 Stock closing prices

As an exploratory step, we first applied the clustering algorithms indicated in Section 4 to the set of time series and plotted the resulting clusters side to side as shown in Figure 1. This visualization provides insights into what stocks have common patterns and the relative ability of the different algorithms to generate relevant groups. In this case, Affinity Propagation showed the least ability to generate groups. Spectral clustering, on the other hand, achieved a spreading of the time series across multiple groups in such a way that stocks in the same group have a similar pattern.

Figure 2 shows the impact of the clustering level and the number of clusters on the silhouette coefficient. Generally, the coefficient decreases with the number of clusters. As stocks are broken into more groups, the separation factor b decreases; without a corresponding gain in cohesion a, the $1 - a/b$ coefficient decreases. On the other hand, the silhouette coefficient is relatively insensitive to the clustering level.

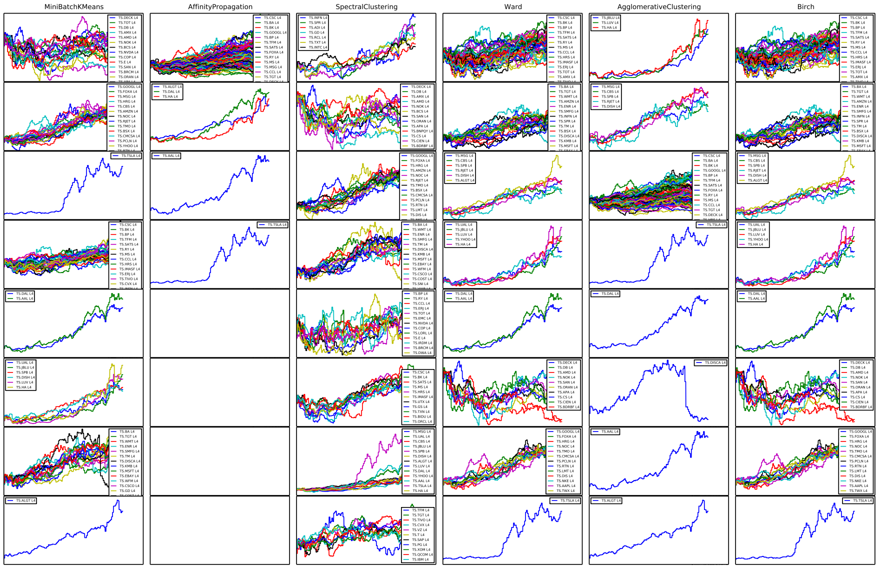


Figure 1: Clustering of stock closing price time series with different algorithms using feature extraction (resolution) level four.

Spectral Clustering, which qualitatively produced relevant clusterings (Figure 1), has a lower coefficient than Mini-Batch K-Means, Ward and Birch. Only when the number of clusters is greater than six, its coefficient is higher than for agglomerative clustering.

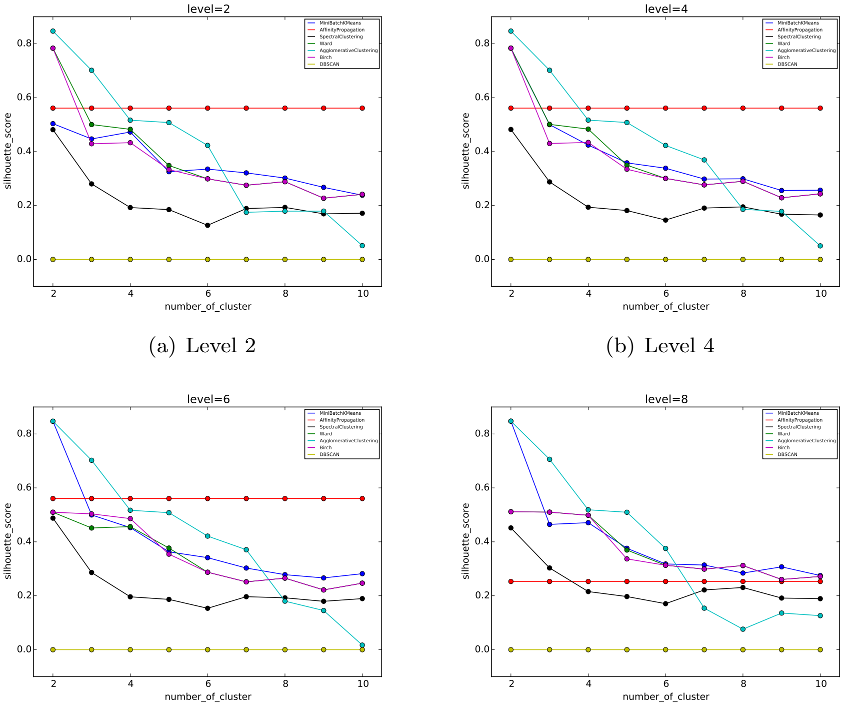
Figure 3 shows again the silhouette coefficient as a function of the clustering level and the number of clusters. As before, the coefficient decreases with the number of clusters, but is relatively insensitive to the clustering level.

4.5 Historic exchange rates

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4.6 Historic earthquake data

On earthquake frequency dataset, we applied a few clustering algorithm to explore relation between clustering and geological distance. Geologically surface of our earth are consisted of some plates, earthquakes are caused by movements of those plates. Therefore, UTM zones might be classified by whether they are



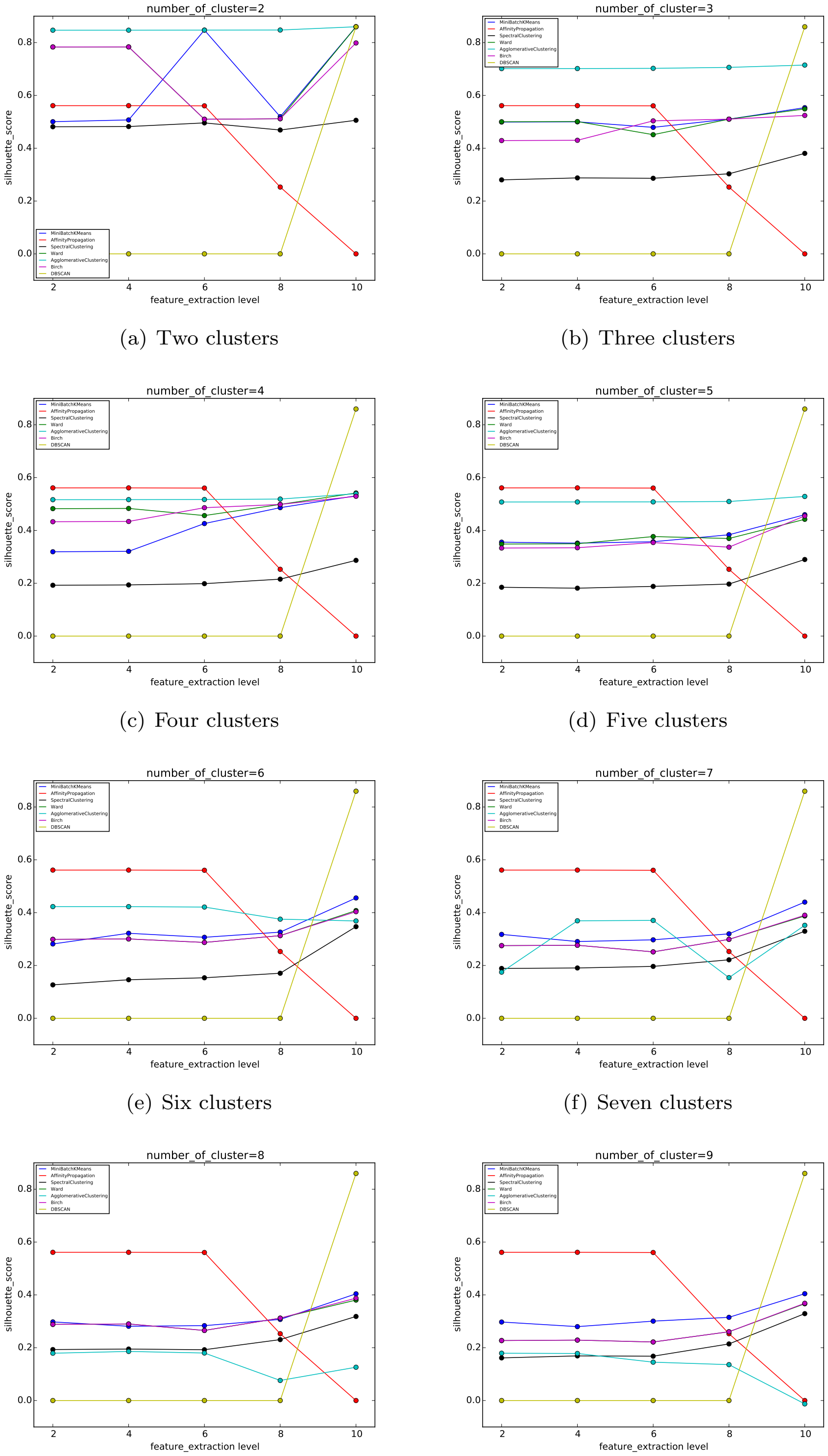
(a) Level 2

(b) Level 4

(c) Level 6

(d) Level 8

Figure 2: Silhouette coefficient for various clustering levels.



(a) Two clusters

(b) Three clusters

(c) Four clusters

(d) Five clusters

(e) Six clusters

(f) Seven clusters

(g) Eight clusters

(h) Nine clusters

Figure 3: Silhouette coefficient for various clustering levels.

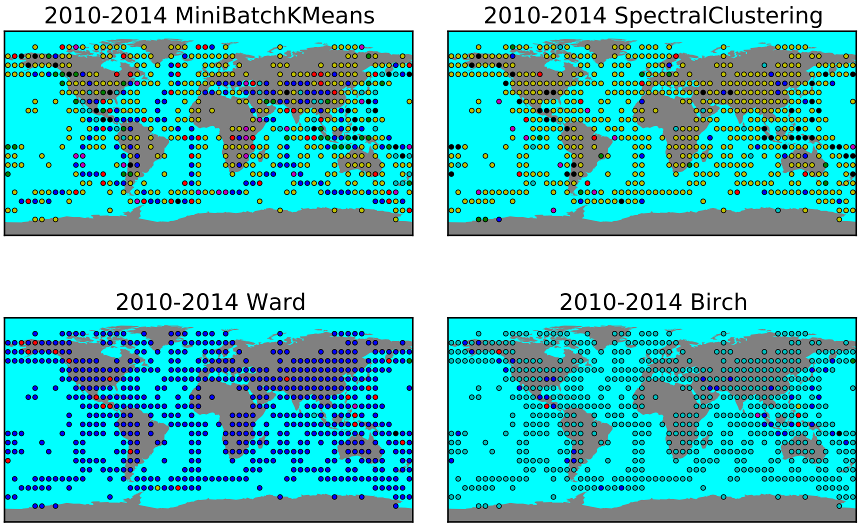


Figure 4: Clustering of earthquake frequency per day with different algorithms using feature extraction (resolution) level six, number of clusters seven.

on the same boundary of plates more than by they distance. Figure 9 shows the UTM-zone in each cluster. Points in the same color indicates, those UTM-zones are in the same cluster and Figure 5 shows tectonic plates of our earth [**?**]. As shown in Figure 2010-2014 Kmeans of 9, we can find some implicit trends that UTM points, on the same boundary in Figure 5, are classified in the same cluster.

Clustering map contains information of not only tectonic plates geology, but also cluster changes over years. Figure 10 shows that clusters in 2010 are migrated in a cluster in 2013. We can assume that earthquake are more related to each other over years by some external causes, such as increasing number of mining on the earth.

Next, we find out what algorithms and parameter setting are fit to earthquake dataset. Figure 7 shows the impact of the clustering level and the number of clusters on the silhouette coefficient. Clustering algorithms are not affected by number of clusters as much as stock price data is affected, except for k-mean clustering. K-means shows drastic changes when the number of cluster is small, since there are geologically a few plates.

Figure 8 shows again the silhouette coefficient as a function of the clustering level and the number of clusters. Unlike the Figure 3 of stock price,

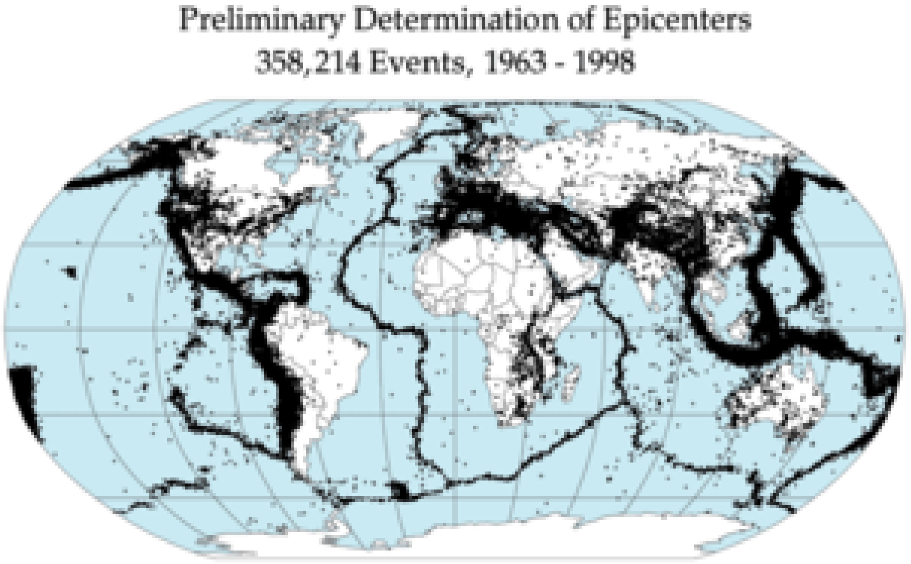

Preliminary Determination of Epicenters
358,214 Events, 1963 - 1998

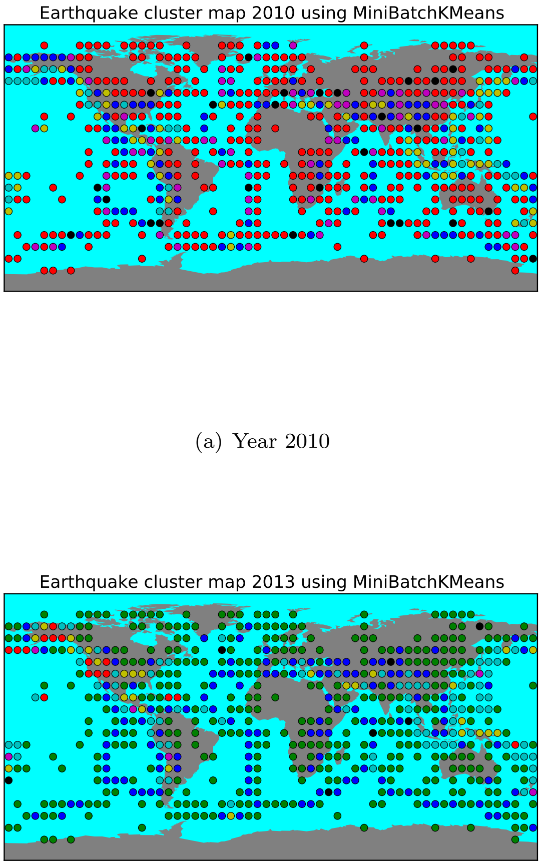
Figure 5: Plate tectonic map of our earth.

earthquake data set is fit to clustering algorithms with high feature extraction level(resolution), since earthquake frequency dataset are zero at most case, and there are earthquake in each UTM-zone for only a few days a month or a year. The higher level of feature extraction is, the more precise information we can pull out of such a trivial changes of earthquake frequency data.

5 Conclusions

This project investigated the use of wavelet transformations as a feature extraction mechanism for the clustering of time series. A framework to evaluate the concept was developed and tested on three different time series data sets: closing prices for stocks, currency exchange rates, and earthquake data.

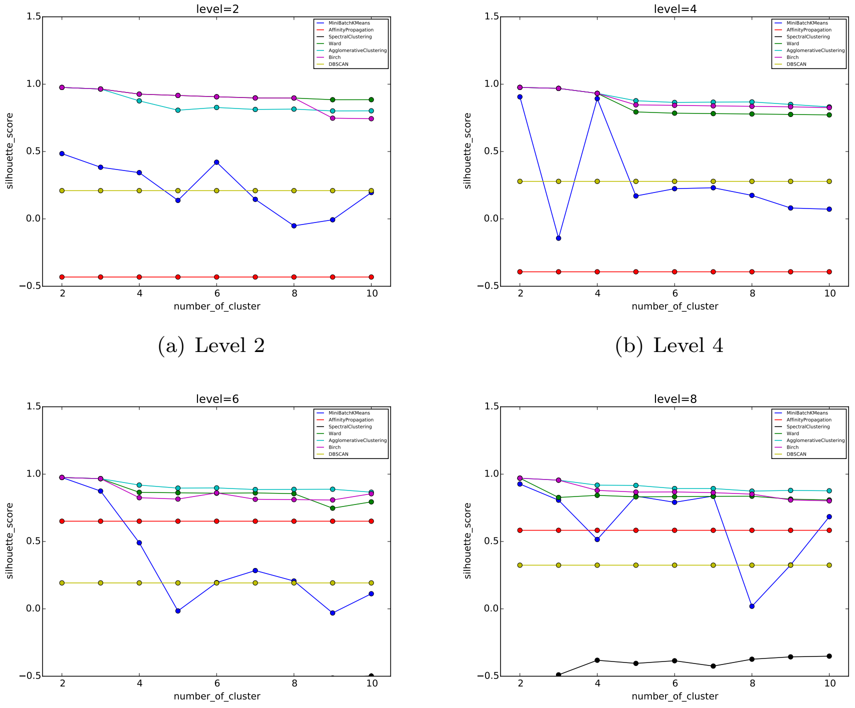
The results of the experimental evaluation suggest that *both* a qualitative assessment as well as the evaluation of a number of measurements, such as the silhouette coefficient, are necessary to understand the data set and the particular combination of parameters for which a clustering algorithm offers the best performance. A premise of the project was that wavelet transformation as a feature extraction mechanism could potentially result in better clusterings. However, the results show that the silhouette coefficient is relatively insensitive to the degree of dimensionality reduction provided by the transformation (clustering level). As part of a process to improve tue quality of the clusterings, compementary performance measurements would need to be included in the evaluation.


Earthquake cluster map 2010 using MiniBatchKMeans
Earthquake cluster map 2013 using MiniBatchKMeans

(a) Year 2010

(b) Year 2013

Figure 6: Clustering of earthquake frequency per day in 2010 vs. 2014 with K-mean using feature extraction (resolution) level six, number of clusters seven.



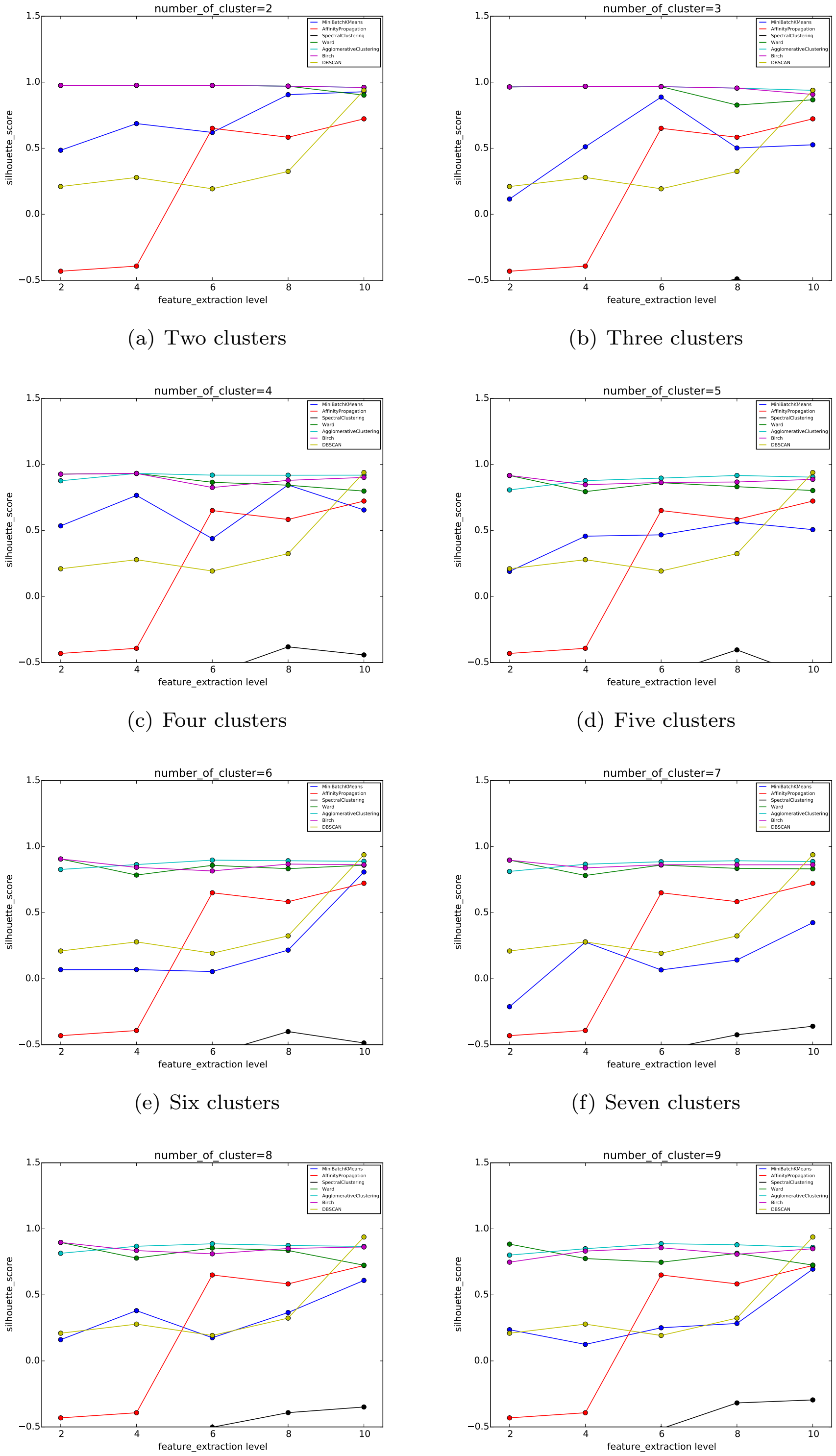
(a) Level 2

(b) Level 4

(c) Level 6

(d) Level 8

Figure 7: Silhouette coefficient for various clustering levels on earthquake frequency dataset.



(a) Two clusters

(b) Three clusters

(c) Four clusters

(d) Five clusters

(e) Six clusters

(f) Seven clusters

(g) Eight clusters

(h) Nine clusters

Figure 8: Silhouette coefficient for various clustering levels on earthquake frequency dataset.

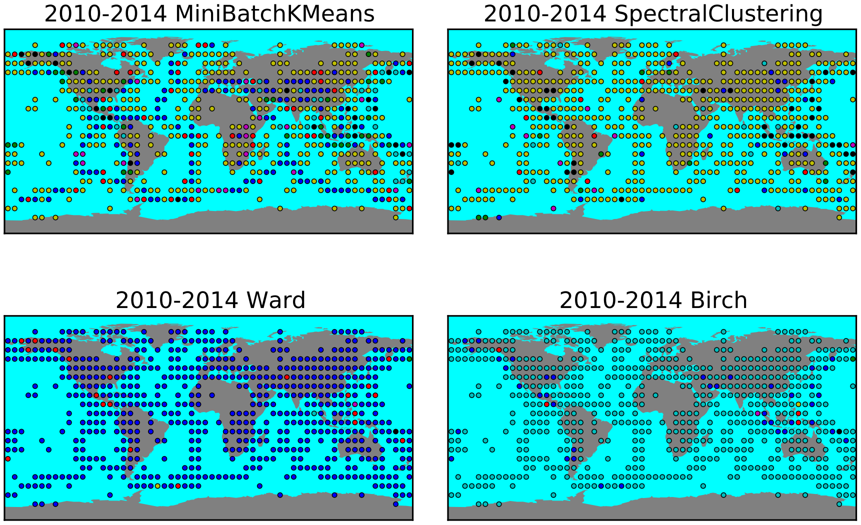
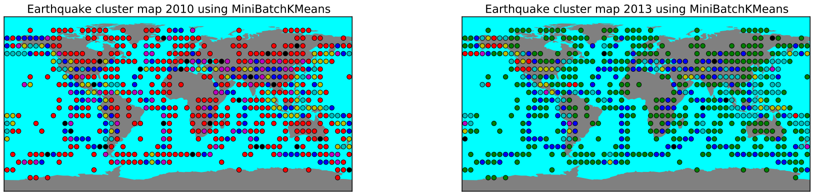


Figure 9: Clustering of earthquake frequency per day with different algorithms using feature extraction (resolution) level six, number of clusters seven.


Earthquake cluster map 2010 using MiniBatchKMeans
Earthquake cluster map 2013 using MiniBatchKMeans

(a) Year 2010

(b) Year 2013

Figure 10: Clustering of earthquake frequency per day in 2010 vs. 2014 with K-mean using feature extraction (resolution) level six, number of clusters seven.

Appendix A: Stock symbols

YHOO, GOOGL, AAPL, MSFT, BIDU, IBM, EBAY, ORCL, CSCO, SAP, VZ, T, CMCSA, AMX, QCOM, NOK, AMZN, WMT, COST, TGT, CVX, TOT, BP, XOM, E, COP, APA, GS, MS, BK, CS, SMFG, DB, RY, CS, BCS, SAN, BNPQY, NKE, DECK, PCLN, EMC, INTC, AMD, NVDA, TXN, BRCM, ADI, WFM, TFM, INFN, CIEN, CSC, TMO, BSX, TIVO, DISH, SATS, LORL, ORAN, IMASF, IRDM, HRS, GD, BA, LMT, NOC, RTN, TXT, ERJ, UTX, SPR, BDRBF, AAL, DAL, HA, UAL, LUV, JBLU, ALGT, RJET, RCL, CCL, DIS, CBS, FOXA, QVCA, DWA, VIAB, TM, TWX, DISCA, SNI, MSG, PG, ENR, HRG, SPB, KMB, TSLA.

References

[1] K. Beyen, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, pp. 217-235, 1999.

[2] H. Zhang, T. B. Ho, Y. Zhang, M.-S. Lin. Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform. In *Informatica*, Volume 30, pp. 305-319, 2006.

[3] C. K. Chui. *An Introduction to Wavelets.* Academic Press, San Diego, 1992.

[4] I. Daubechies. Ten Lectures on Wavelets. *SIAM*, Philadelphia, PA, 1992.

[5] J. Morlet, G. Arens, E. Fourgeau and D. Giard. Wave propagation and sampling theory, Part 1: Complex signal land scattering in multilayer media. *Journal of Geophysics,* 47:203-221, 1982.

[6] J. M. Combes, A. Grossman and P. Tchamitchian, editors. *Wavelets, Time-Frequency Methods and Phase Space.* Springer-Verlag, Berlin 1989.

[7] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics,* 41:909-996, 1988.

[8] C. Sidney Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer,* Prentice Hall, NJ 1998.

[9] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining,* Pearson, First Edition, 2005.

[10] Federal Reserve Economic Data (FRED) service, *https://research.stlouisfed.org/fred2/.* Accessed: May 27th 2015.

[11] SciKit Learn *http://scikit-learn.org/stable/.* Accessed: May 27th 2015.