Indian Institute of Technology Indore
Department of Computer Science and Engineering

CS-403
MACHINE LEARNING

Forecasting Financial series using Clustering Methods and Support Vector Regression

Two-Stage Model for Forecasting Financial Time Series

PRESENTED BY

- Rahul Kumar
- Vivek Bhushan
- Harsh Vardhan Agrawal

Introduction

Forecasting financial series using clustering and SVR

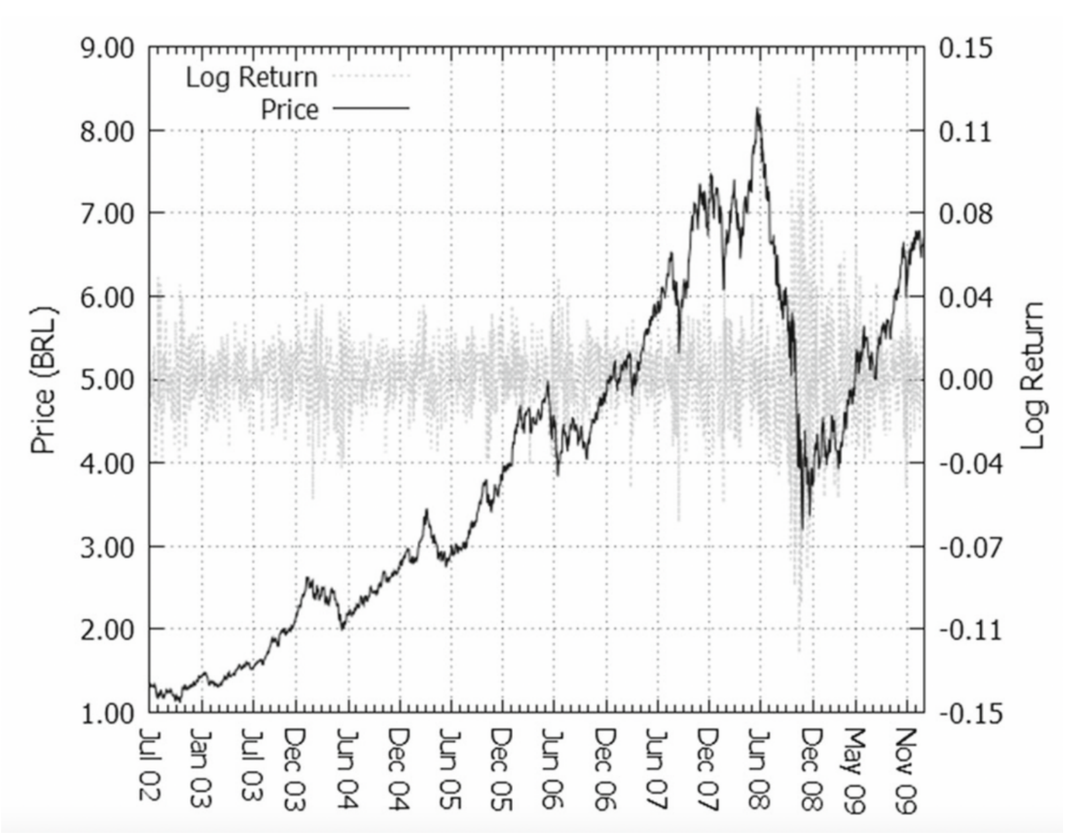
- A time series S is a sequence of values (or observations) of an event where the values have a defined periodicity.
- Financial time series are characterized by noise and non-stationarity, making modeling and forecasting more challenging.
- The creation of forecasting models for time series began with linear statistical models. These models had restrictions.
- Nonlinear models, such as artificial neural models and learning machines, have been developed to treat the non-linearities of the series better.
- Machine learning models, constructed exclusively from the series' values, are less susceptible to noise and have been widely used for financial time series analysis and forecasting. SVMs are among the most widely used machine learning models for forecasting time-series values.

Data Preprocessing

- A time series is defined as a sequence of observations obtained in chronological order.
- They present a defined type of stochastic behavior.
- The **continuous compound return** value is considered for analyzing financial time series as it carries desirable features for analyzing asset price behavior.

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1})$$

- The dataset used consists of the daily price values of an equity fund managed by a Brazilian bank. It is a time series data that contains the closing price at the end of each trading day.



Making Correlations and Finding Features

- The continuous compound return series is weakly stationary and doesn't show trend or scale components.

- On weakly stationary series, both the mean of r_t (return at time t) and the covariance between r_t and r_{t-l} are time-invariant where l is the lag.

- By this assumption, we can estimate the autocovariance function by computing the lag-l autocovariance $\gamma_l = Cov(r_t, r_{t-l})$ and the autocorrelation function by computing the lag-l autocorrelation $\rho_l = Cor(r_t, r_{t-l})$

Making Correlations and Finding Features

$$\gamma_l = Cov(r_t, r_{t-l}) \qquad \rho_l = Cor(r_t, r_{t-l})$$

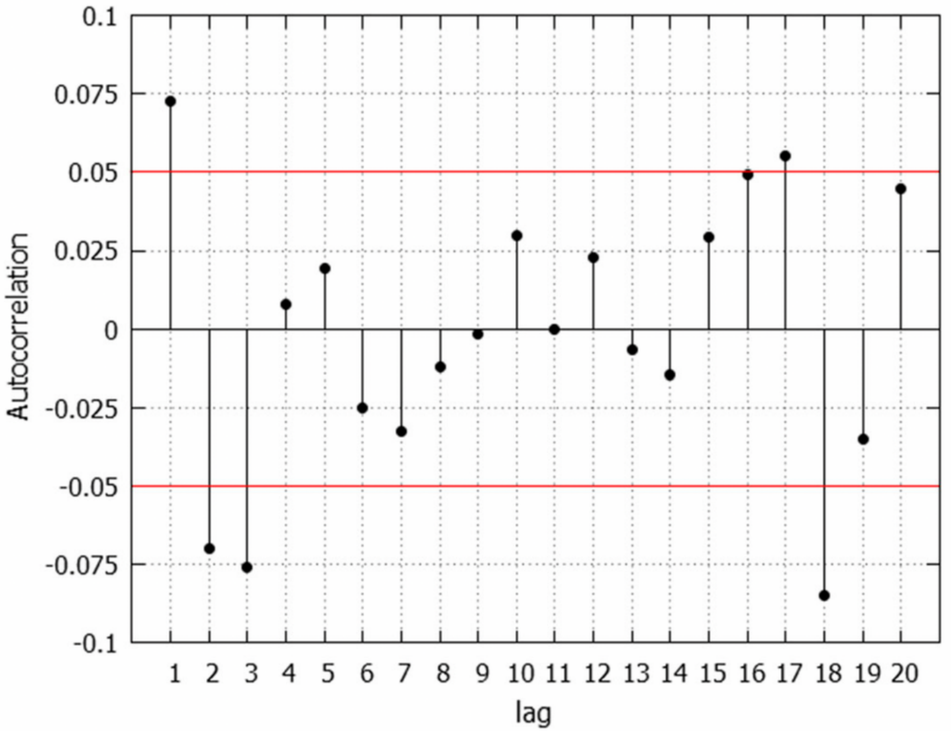
$$\rho_\ell = \frac{\text{Cov}(r_t, r_{t-\ell})}{\sqrt{\text{Var}(r_t)\text{Var}(r_{t-\ell})}} = \frac{\text{Cov}(r_t, r_{t-\ell})}{\text{Var}(r_t)} = \frac{\gamma_\ell}{\gamma_0},$$

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{T}(r_t - \overline{r})(r_{t-1} - \overline{r})}{\sum_{t=1}^{T}(r_t - \overline{r})^2}.$$

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^{T}(r_t - \overline{r})(r_{t-\ell} - \overline{r})}{\sum_{t=1}^{T}(r_t - \overline{r})^2}, \quad 0 \leq \ell < T - 1.$$

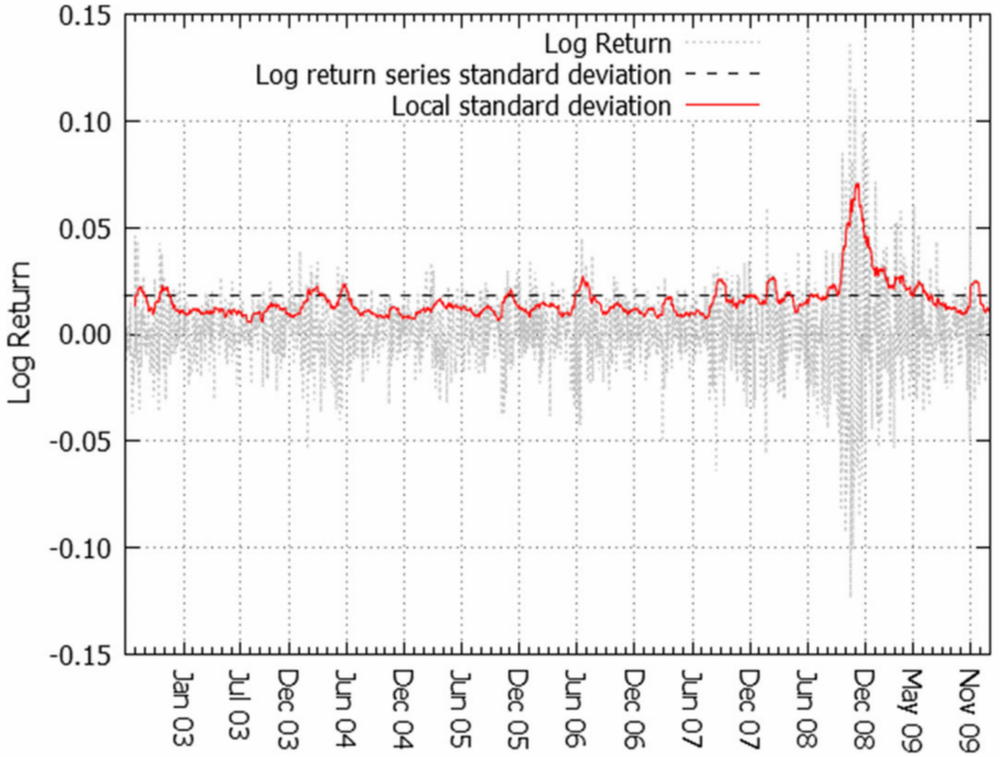
Making Correlations and Finding Features

- The statistical method of autocorrelation is applied to the return series.
- Using these values, an auto-correlogram was plotted to depict the relationship of the time series from its lagged version. Each stem refers to the estimated lag-l autocorrelation $\hat{\rho}_l$.
- Further, to choose which values are significant, a significance of 5% was considered.



Analysis of Time Series Volatility

- In financial time series, uptrend periods usually present lower volatility than downtrend periods.
- A rising price (bullish) market usually moves slower than a declining price (bearish) market, impacting the security's return and volatility
- Return series can be divided into periods of high and low volatility based on available data.
- The volatility of a value "s_i" of the series is given by comparing the standard deviation of the return series with its local standard deviation (standard deviation of the range of the twenty values prior to "s_i").



Final Feature Matrix

- The condition attributes are formed by the three past values — r_{t+f-1}, r_{t+f-2}, and r_{t+f-3} — correlated with the value to be forecast, and by the volatility of the value to be forecast.
- Volatility is a binary value where the values 0 and 1 mean low and high volatility, respectively.
- The decision attribute is formed by the value to be forecast, i.e., future values r_{t+f} of the return time series.

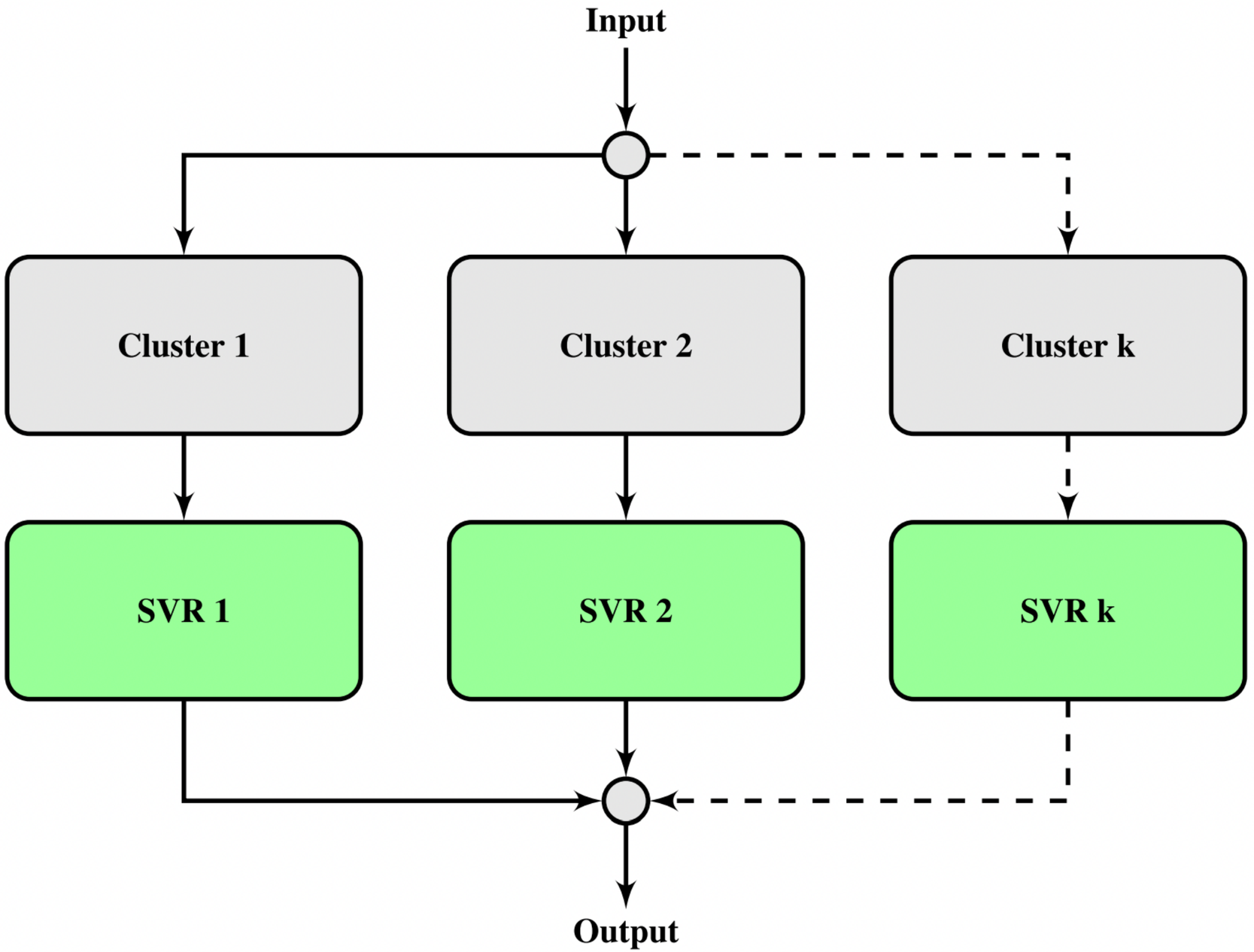
Pattern (condition attributes)				Decision attribute
r_{t+f-3}	r_{t+f-2}	r_{t+f-1}	Vol	r_{t+f}

- Define the Problem
- Collecting Raw Data
- Feature Geneartion
- Model Building & Training
- Experimentation

Proposed Model Architecture

- Clusters are formed through the execution of either the K-Means method or the Fuzzy C-Means method on the patterns of the decision table.

- The next training phase involves training the corresponding SVR on the patterns inside the cluster for each cluster created in the previous phase.



How Predictions are made

- When an input pattern is presented to the model, its first stage determines the membership of the pattern in relation to the k clusters of the model.
- The SVRs in the second stage perform the regression, using the knowledge previously acquired in their respective training, to predict the next value of the return series.
- Now because of multiple SVRs, there are two methods for prediction.

Using Single SVR

- A single SVR, relating to the cluster to which the input pattern has the highest membership, is responsible for the forecast.
- This approach can be used by hard clustering as well as soft clustering methods

Using All SVRs

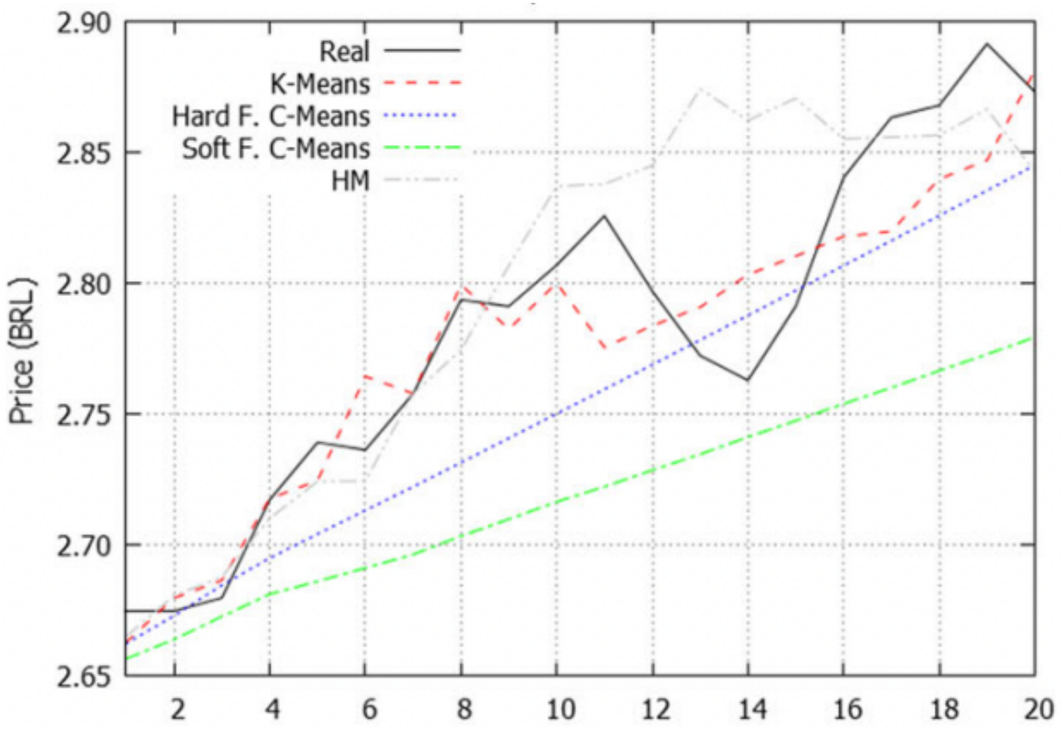
- All SVRs are responsible for the forecast, each one yielding its own forecast value. The degrees of membership of the input pattern relating to the clusters are used as weights.
- Only soft clustering methods can be used in this approach.

Experimentation

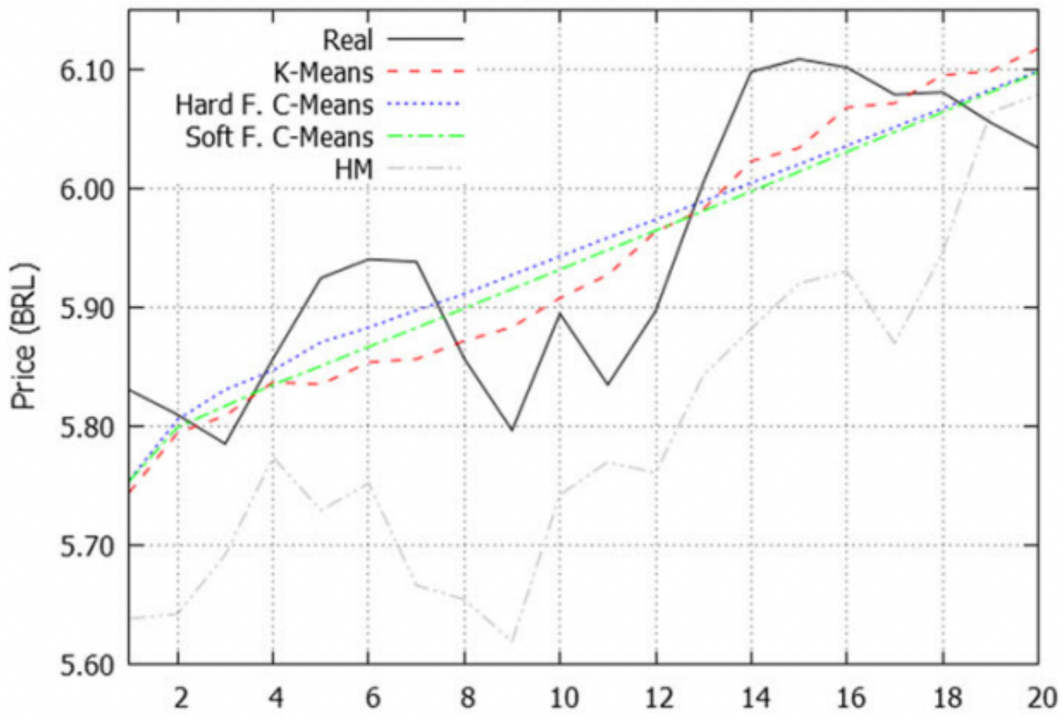
- Six experiments are performed aimed at forecasting a period of time, comprising twenty consecutive days, of the return series.
- With the K-Means method, only the single SVR approach was used, for it is a hard clustering method.
- The SVRs used the radial basis function as the kernel function, since this produced the best results reported in a previous research.

Experiment	Time period		Volatility	Trend
	Start	End		
1	2004-11-22	2004-12-17	Low	Uptrend
2	2007-05-28	2007-06-25	Low	Uptrend
3	2009-09-11	2009-10-08	Low	Uptrend
4	2006-05-19	2006-06-16	High	Correction
5	2008-01-18	2008-02-18	High	Horizontal
6	2009-10-14	2009-11-11	High	Horizontal

Results

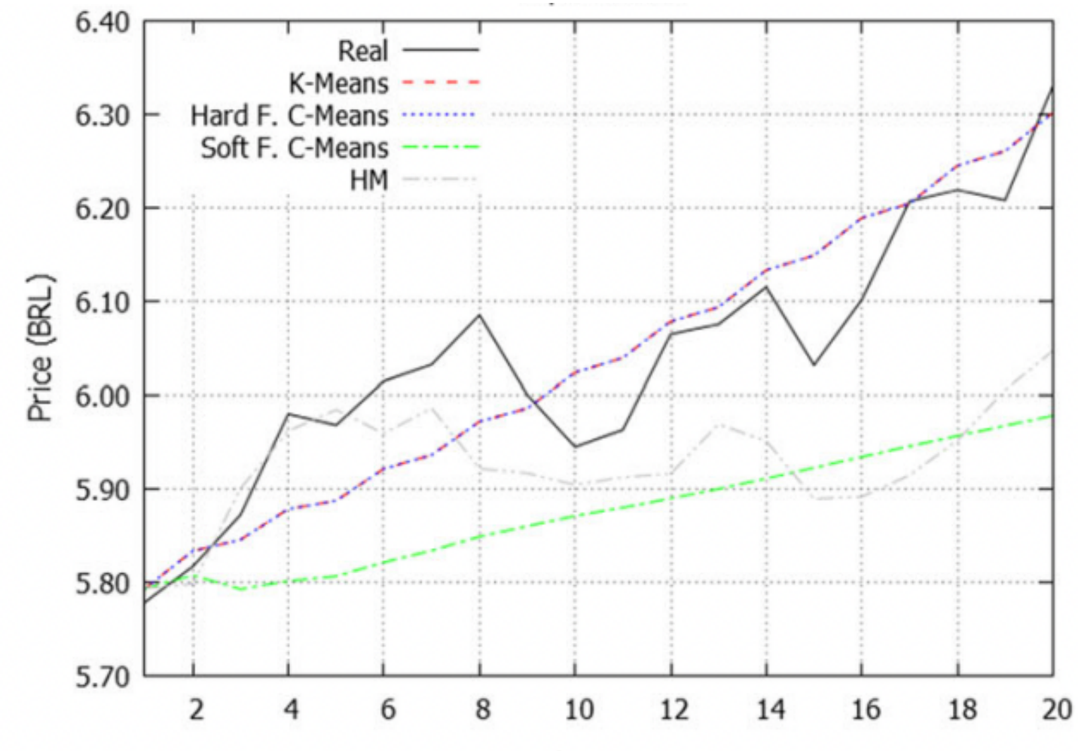


Experiment 1

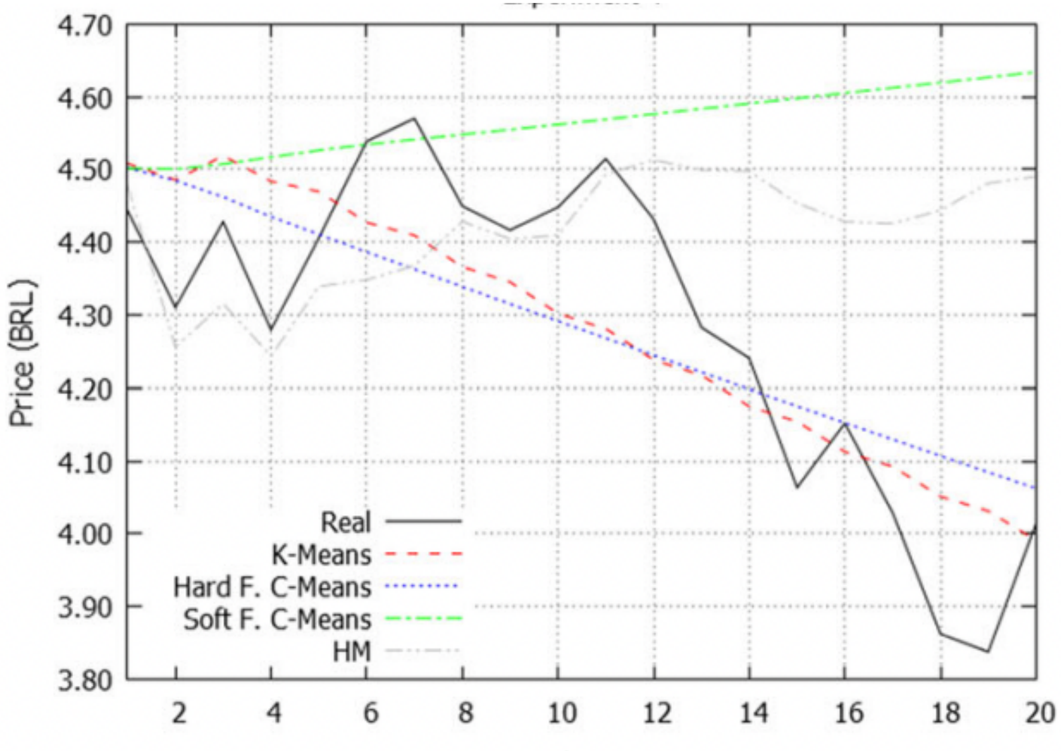


Experiment 2

Results

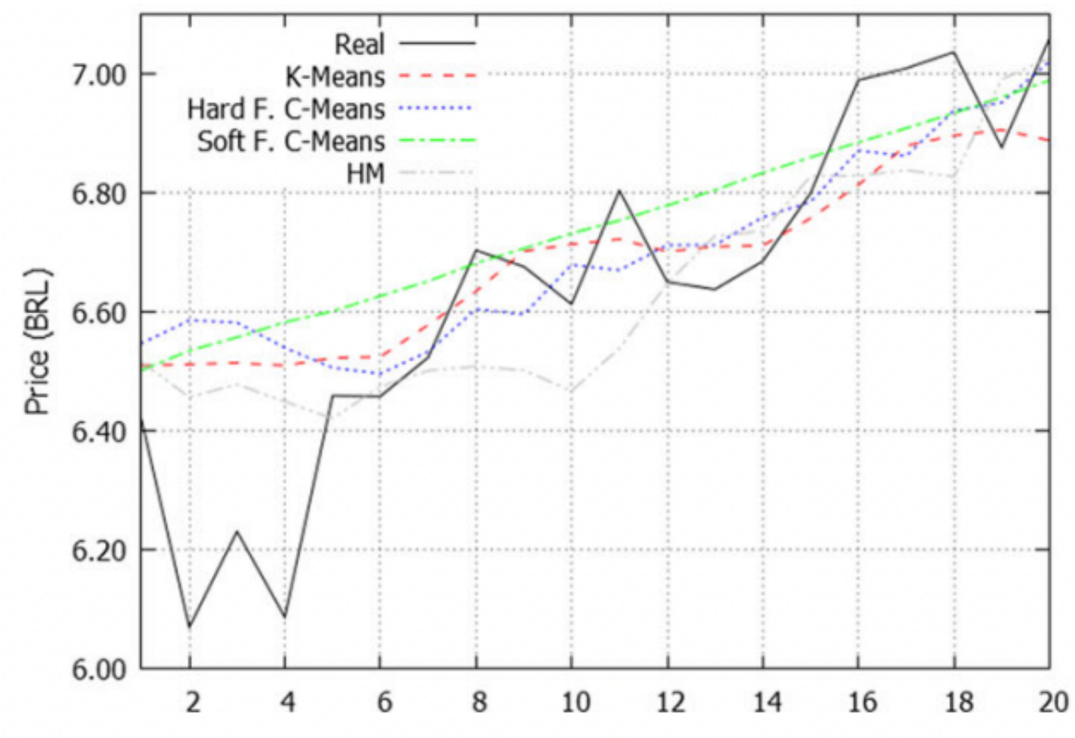


Experiment 3

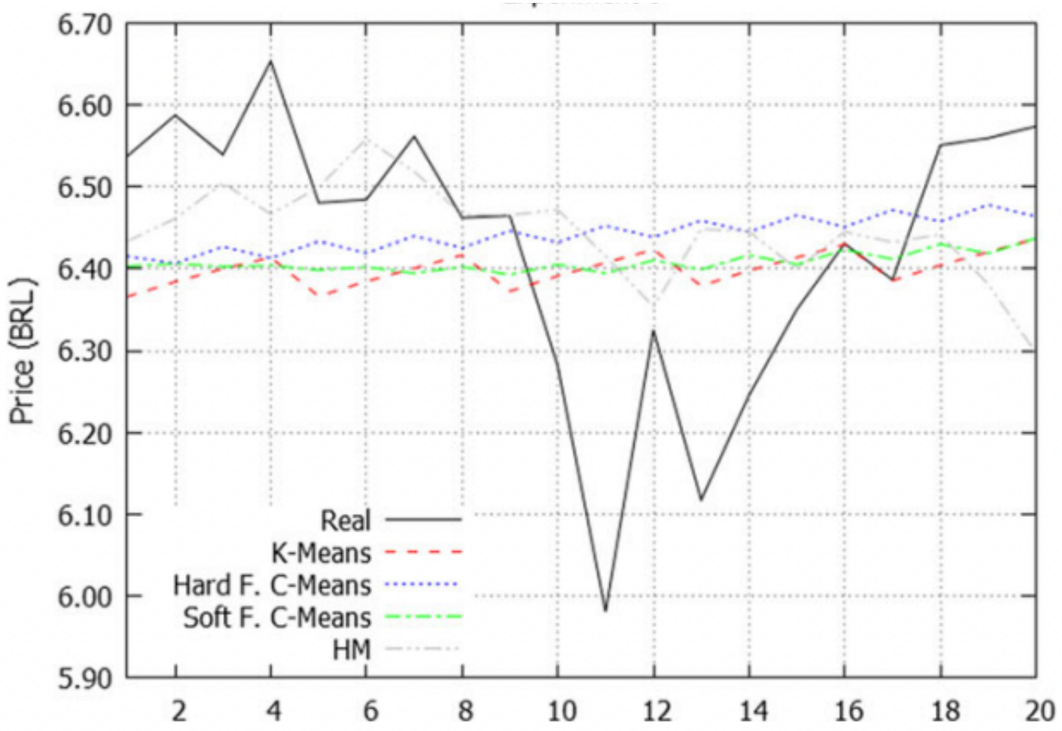


Experiment 4

Results



Experiment 5



Experiment 6

Results

Experiment	K-Means	Hard Fuzzy C-Means	Soft Fuzzy C-Means	HM
1	0.024	0.037	0.072	0.041
2	0.061	0.066	0.066	0.165
3	0.066	0.066	0.186	0.149
4	0.132	0.145	0.375	0.282
5	0.174	0.190	0.196	0.179
6	0.168	0.174	0.164	0.169
Mean	0.104	0.113	0.176	0.164

Root Mean Square Error(RMSE) value in each experiment

References

Code Link:

- https://colab.research.google.com/drive/16-3WYPJrc9M9kiEjKN6JeEFKN39Kl8JU#scrollTo=od24m3POZcvf

References:

- https://link.springer.com/article/10.1007/s10462-018-9663-x
- https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/0/6796/files/2017/03/analysis-of-financial-time-series-copy-2ffgm3v.pdf

Thank You

Appendix

Testing ACF

- If $\{r_t\}$ is an independent and identically distributed sequence satisfying $E(r_t^2) < \infty$, then $\hat{\rho}_\ell$ is asymptotically normal with mean zero and variance 1/T for any fixed positive integer ℓ

- This result can be used in practice to test the null hypothesis $H_0 : \rho_\ell = 0$ versus the alternative hypothesis $H_a : \rho_\ell \neq 0$.

- The test statistic is:

$$t\text{-ratio} = \frac{\hat{\rho}_\ell}{\sqrt{(1 + 2\sum_{i=1}^{\ell-1} \hat{\rho}_i^2)/T}}.$$

- If $\{r_t\}$ is a stationary Gaussian series satisfying $\rho_j = 0$ for $j > \ell$, the t-ratio is asymptotically distributed as a standard normal random variable. Hence, the decision rule of the test is to reject H_0 if |t-ratio| > $Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the 100(1 − α/2)th percentile of the standard normal distribution.

Fuzzy C-Means

- The Fuzzy C-Means method allows partial membership of objects to clusters.
- If u_{ij} is the membership of object x_j in cluster i, its value can vary from 0, which means no membership, up to 1, indicating total membership.
- It is classified in the soft clustering method because it allows partial membership.
- The degree of membership u_{ij} can be calculated using the given formulae:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}^2}{d_{il}^2} \right)^{\frac{1}{m-1}}}$$

- d_{ij} here represents the distance between the object and the centroid of the cluster.
- And m represents the fuzying parameter if m=1 it behaves like Hard Clustering and when m>1 the membership is partial.
- The basic rules of Fuzzy-C means clustering are.
  - Empty clusters are not allowed.
  - The sum of the object's memberships in the clusters must be equal to 1.