

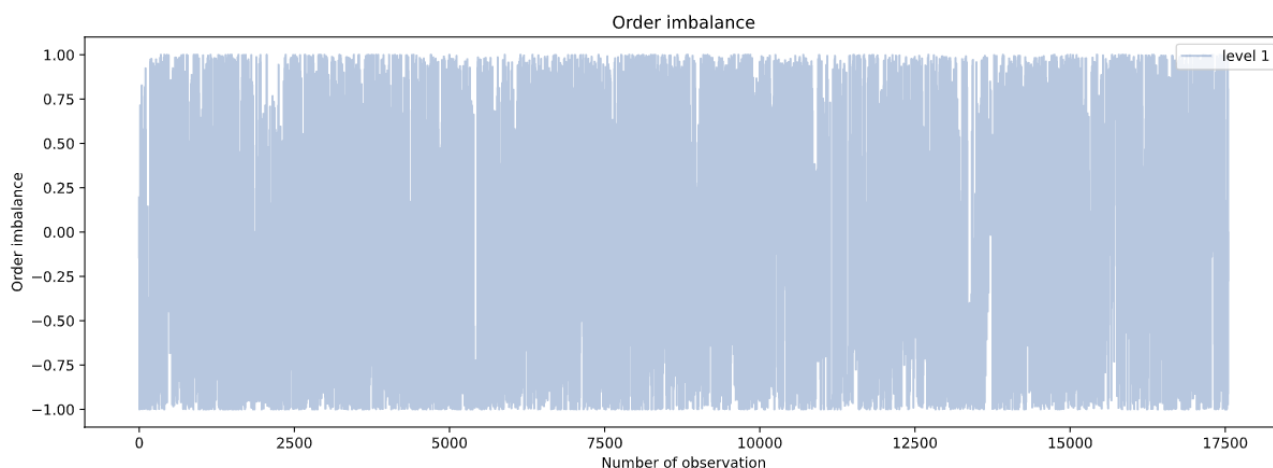
Articles by financial experts

I have searched for similar articles from financial experts on the topic, but there are only few of them. In particular, I have not found any where DBSCAN or KPCA are used. What concerns application of clustering method, I found two articles where authors use Gaussian Mixture model. These are a research by financial services company IHS Markit (*Using a “Human + Machine” Approach to Determine Financial Market Regimes*) and article *Gold and silver manipulation: What can be empirically verified?* by J. A. Batten, B. M. Lucey and M. Peat. IHS Markit experts used clustering for a different purpose – they clustered implied volatility according to the market state (crisis, transition, etc.) it came from. However, they also encountered the problem of how to assess a model, since their data was also unlabeled. They propose two methods: visual inspection and computing silhouette coefficient (the one I wrote about in the previous email). The second article uses Gaussian mixture model to detect manipulation in gold and silver markets. This model consists in finding parameters for a mixture (combination) of normal distributions, so that the estimated density is most likely to have generated the data. The parameters are chosen for each normal distribution in the mixture. These normal distributions are called components. The authors find the best result with 3 components, and they obtain parameters for these components and their weights in the mixture. The component with largest volatility, smallest weight and lowest mean is claimed to mark manipulative trades. When a trade arrives, a Bayes classifier assigns this trade to the component by which it is most likely to be generated. Then the authors separate the component to which most observations are assigned and call it control group of normal trades. They run several statistical tests, comparing statistics of control group and of manipulative component.

Finally, there is a comprehensive overview of manipulation detection methods in various areas (*Statistical Methods for Fighting Financial Crimes* by Sudjianto, Agus, et al.). The authors briefly describe the models and then show their application on some tasks. In particular they used Hidden Markov model for detection of checking account overdraft fraud. They also mention PCA and density-based outlier detection, similar to DBSCAN.

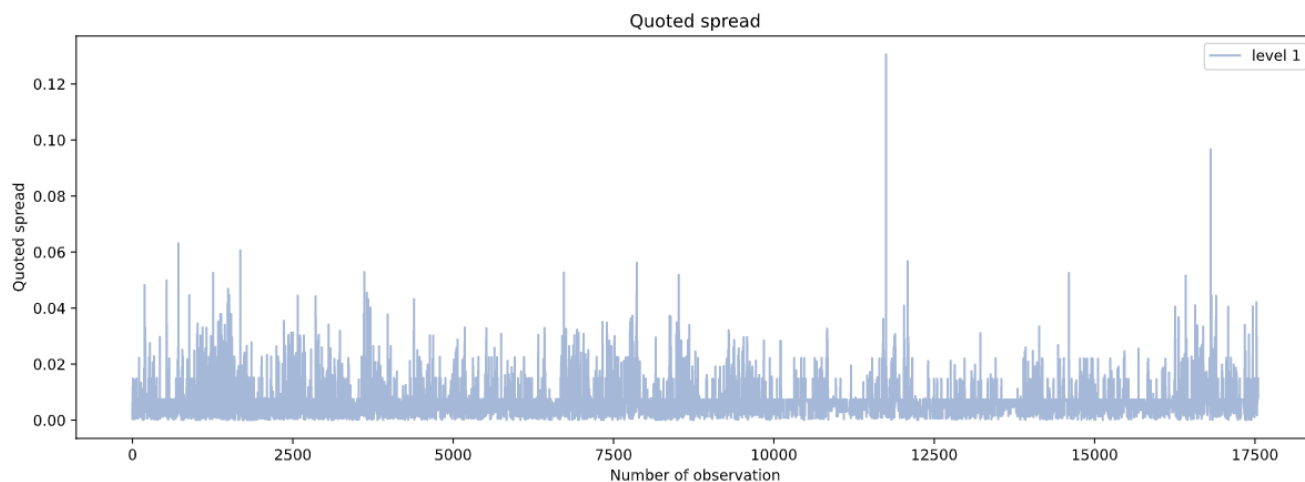
Features

$$\text{Order imbalance} = \frac{\text{bid volume} - \text{ask volume}}{\text{bid volume} + \text{ask volume}}$$

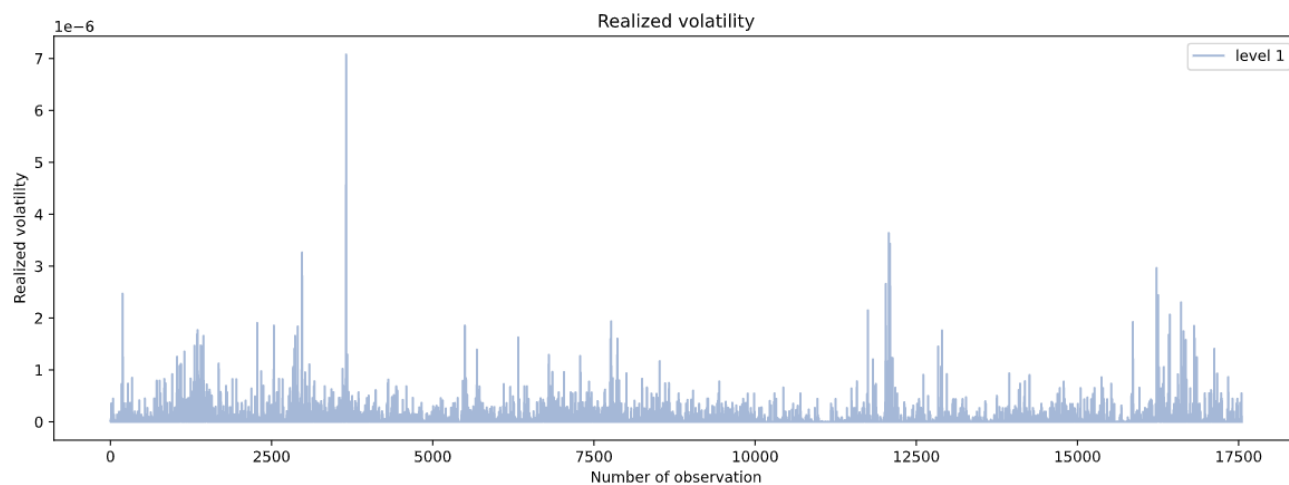


Quoted spread

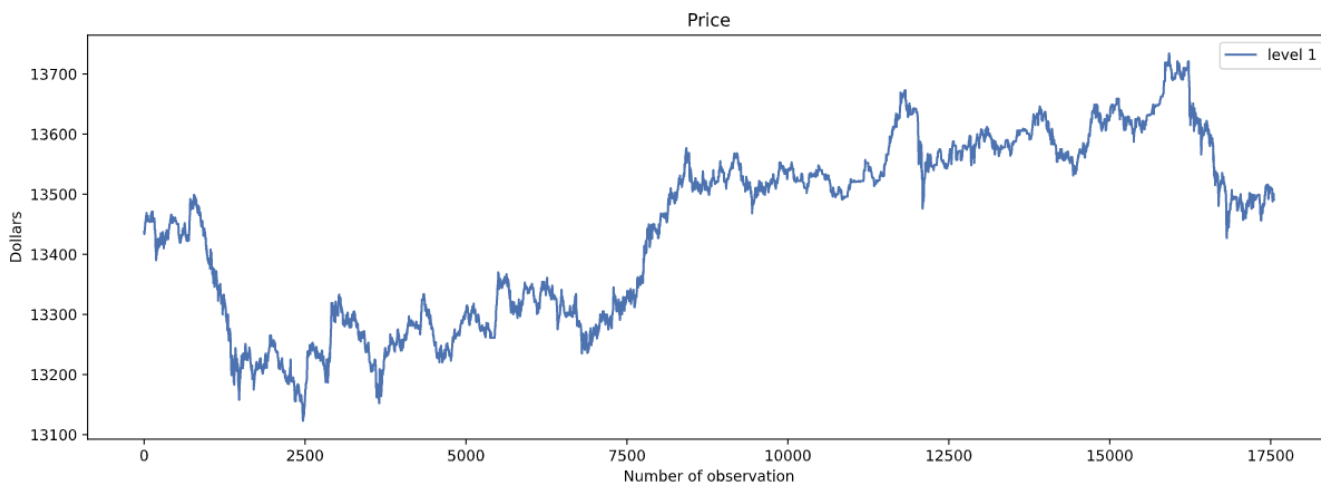
$$\text{Quoted Spread} = \frac{\text{ask} - \text{bid}}{\text{midpoint}} \times 100$$



Realized volatility – proxied by squared returns

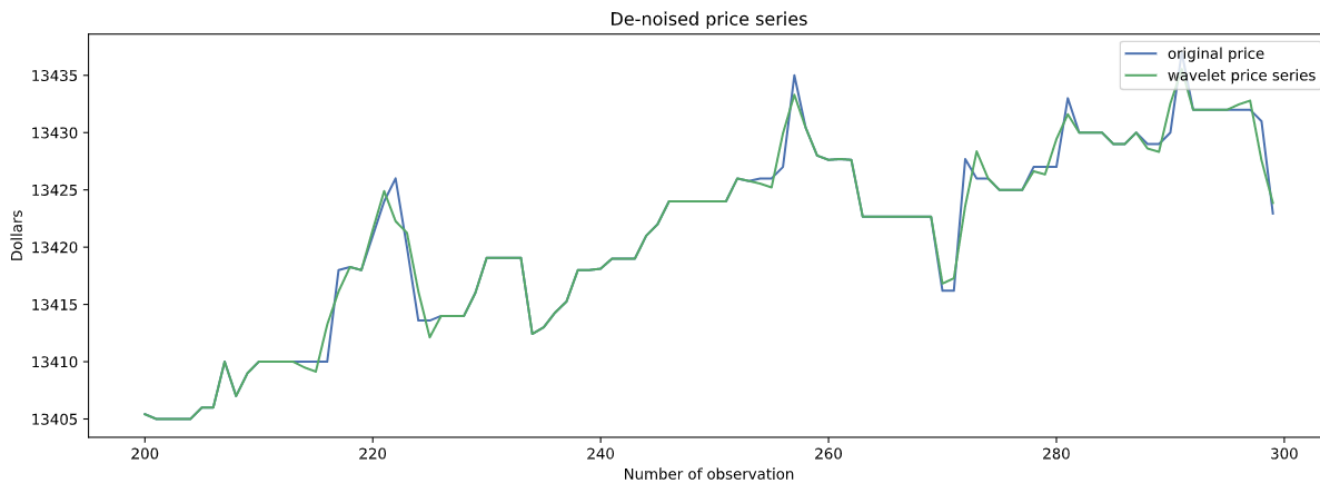


Price

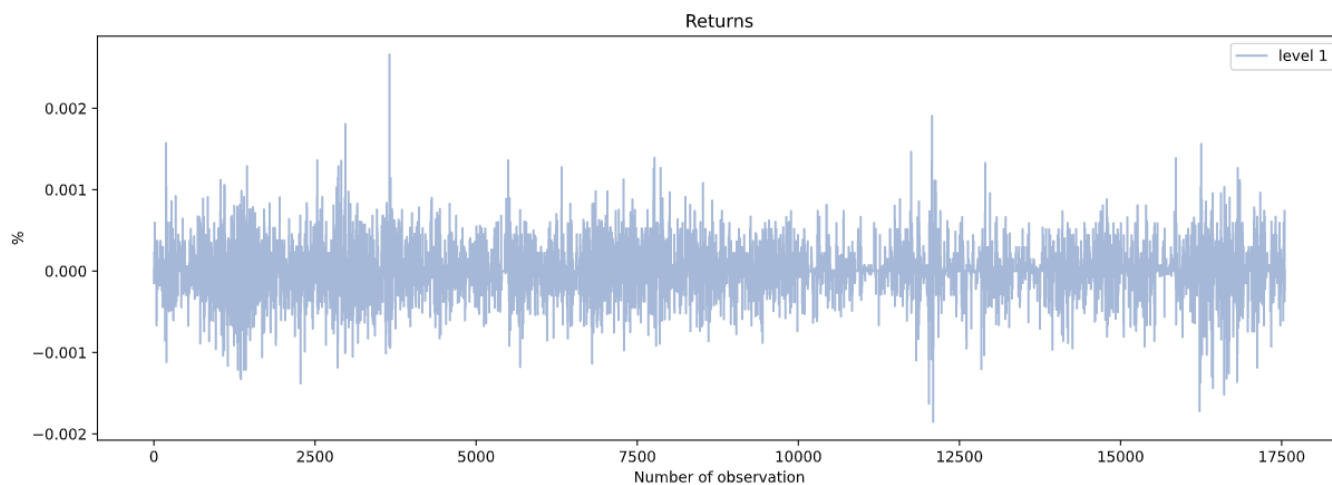


De-noised price (with discrete wavelet transform)

De-noised time series constructed with wavelets is smoother than the original one:



Returns of original price



Returns of de-noised price

