MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Evaluation of Volatility Forecasting Models“

verfasst von / submitted by

Bernhard Kober, BA BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 951
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von / Supervisor:	ao. Univ.-Prof. Dipl.-Ing. Dr. Erhard Reschenhofer

Contents

List of Figures

1 Introduction

Time series analysis is a branch of statistics, that has a close relationship to financial mathematics, especially when it comes to application in stock markets. The mathematical foundation is very profound, while the methods for volatility forecasting are progressing every year, bringing forth different approaches and new models, always in consideration of the newest available data. One of those new approaches was made by Bollerslev at al. [BPQ16], in an attempt to exploit the errors. In their released article “Exploiting the errors: A simple approach for improved volatility forecasting”, Bollerslev et al. [BPQ16] introduced a new model for volatility forecasting that can be viewed as a composition of earlier models. This new model was then compared to a variety of different models, where their performances were examined thoroughly. However, understanding the behavior of forecasting models, or even volatility itself, is by no means a trivial thing. In his Book “Models Behaving Badly”, Emanuel Derman [Der11] gives insight about his viewpoint of financial models and expresses his critique about models being assigned a level of reliability that they simply cannot offer, as they are confused with being a description of reality, whilst actually providing just a simplified view of complex processes, which otherwise could not be comprehended. Goldstein and Taleb [GT07] even claim, that “We don’t quite know what we are talking about when we talk about volatility”, which is also the title of their paper published in 2007. They conducted an experiment, asking professionals in investment management companies, ivy league graduate students and professionals working in a major bank, to calculate – given several assumptions – the daily volatility and the yearly volatility. The clear majority arrived at the same – but wrong – answer, confusing standard deviation with mean absolute deviation (out of 87 responses, only three showed the correct solution). Considering the influence that stock markets have on the whole world, it is only reasonable to let the critical voices be heard – giving up on financial models and be groping in the dark would not be a good solution though. The problem of understanding forecasting models and to reasonably assess their reliability might not be limited to mathematical and statistical properties, however, the least that can be done, is to take a closer look at and to widen the exposure of new concepts, so that we can draw from a more substantial range of observations.

Thus, this papers purpose is to reproduce and extend the evaluation that was performed in Bollerslev et al. [BPQ16]. The mathematical principals and definitions required to understand the following models and analysis will only be discussed briefly in Chapter 2, since a superficial introduction to the basic concepts should suffice to follow the explications - further in-depth literature is readily available online. In Chapter 3, all models taking part in the evaluation will be introduced. In Chapter 4 we will take a look at the data that was used for the analysis, followed by the eventual evaluation and discussion in Chapter 5. Finally, Chapter 6 concludes with the findings of this paper and prospects for the domain of evaluating volatility forecasting models.

2 Definitions

Consider the price process Pt for an arbitrary stock, then

$$d\log(P_t) = \mu_t dt + \sigma_t dW_t$$

denotes a standard Brownian motion, whereas mu is the drift process and sigma is the instantaneous volatility process in this stochastic differential equation. We basically assume this structure holds for all stocks used for our forecasting models. What we actually want to forecast would be the (daily) latent Integrated Variance

$$IV_t = \int_{t-1}^{t} \sigma_s^2 \, ds$$

which, unfortunately, is not directly observable. However, since we have a consistent estimator in the Realized Variance

$$RV_t \equiv \sum_{i=1}^{M} r_{t,i}^2$$

as $\Delta \rightarrow 0$, where $M = 1/\Delta$, we can focus on forecasting the Realized Variance instead, as it is easy to observe the returns $r_{t,i} = \log(P_{t-1+i\Delta}) - \log(P_{t-1+(i-1)\Delta})$. Besides the Realized Variance, there is another variable that is essential to the models proposed by Bollerslev et al. [BPQ16], the Realized Quarticity – abbreviated as RQ. To understand what the Realized Quarticity is, we again consider the Realized Variance and the Integrated Variance. Let η_t be the estimation error of the daily Realized Variance

$$RV_t = IV_t + \eta_t$$

Then η_t follows a mixed normal distribution with mean 0 and Variance $2\Delta IQ_t$. IQ denotes the Integrated Quarticity

$$IQ_t = \int_{t-1}^{t} \sigma_s^4 \, ds$$

which, just like the Integrated Variance, is not observable. There is, though, a consistent estimator in the Realized Quarticity

$$RQ_t \equiv \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4$$

which, again, is a function of the daily returns.

With the RV being an estimator for the daily volatility and RQ being an estimator for the variance of the estimation error, it is now clear what Bollerslev et al. [BPQ16] suggested by exploiting the errors. By implementing a term for the estimation error into the model, potential forecasting errors from unusual values in the RV are expected to be absorbed by the RQ term, since unusual values in the RV should theoretically and mathematically be accompanied by unusual values in RQ.

3 Models

As already mentioned, there are several models that are to be compared in this paper. Basically, all of them are based on a simple autoregressive model of first order, extended in different ways to attempt to deal with the characteristics of the data. Though these models look quite similar, each of them shall be introduced and explained separately to account for better readability in the following chapters.

3.1 ARQ

$$RV_t = \beta_0 + \left(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}\right) RV_{t-1} + \mu_t$$

The ARQ model is a straight forward extension of the AR(1) model. In adding only the cross-term $\beta_{1Q} RQ_{t-1}^{1/2} RV_{t-1}$ to the right hand side of $RV_t = \beta_0 + \beta_1 RV_{t-1} + \mu_t$, we are getting the above expression as a result. The goal of letting the coefficient of RV_{t-1} vary with the Realized Quarticity, is to let days with large error measurement and little information have less impact on forecasts. This property only holds though, if β_{1Q} is negative. Meanwhile, in the extreme (but unlikely) case of constant Realized Quarticity, this model is again reduced to an AR(1) model.

3.2 HAR

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \mu_t$$

The HAR-RV model – or simply HAR model – by Corsi [Cor09] takes a different approach. The aim here is to create a model that is capable of dealing with long-run dependencies while retaining the simplicity of standard AR models. We find a new expression in this model, which in general form reads as $RV_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} RV_{t-i}$ where $j \leq h$. Consequently, the model is constituted by the addition of the daily RV, weekly RV and monthly RV with respect to a given day t-1. Interestingly, the HAR-RV model appears to be the most popular model for daily Realized volatility forecasting (see [BPQ16], page 4).

3.3 HARQ-F

$$RV_t = \beta_0 + \left(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}\right) RV_{t-1} + \left(\beta_2 + \beta_{2Q} RQ_{t-1|t-5}^{1/2}\right) RV_{t-1|t-5} + \left(\beta_3 + \beta_{3Q} RQ_{t-1|t-22}^{1/2}\right) RV_{t-1|t-22} + \mu_t$$

This model essentially is a combination of the ARQ and the HAR model. Since the HAR model is quite popular, it is reasonable to try and improve it in the same sense in which the ARQ model is an improvement to the AR model. Thus, the coefficients of the weekly RV and the monthly RV are supposed to depend on the weekly and monthly RQ, respectively. Like $RV_{t-j|t-h}$ is used to calculate the weekly and monthly RV, the generic form for the weekly and monthly RQ similarly is $RQ_{t-j|t-h} = \frac{1}{h+1-j} \sum_{i=j}^{h} RQ_{t-i}$ where $j \leq h$.

3.4 HARQ

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})\, RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \mu_t$$

Since the RQ component within our models is brought in to absorb impact of misleading spikes in the data - especially single days that are extreme outliers – the monthly and weekly RQ parts appear to be redundant, since single strong deviations average out anyways if we build the mean for the RV over 5 days or 22 days for weekly and monthly RV respectively. Thus, it is justifiable to reduce the full HARQ model (HARQ-F) to a more compact form as written above (HARQ). The coefficients for the weekly and monthly RV no longer vary with the RQ, while the coefficient for the daily RV still does.

Both, the HARQ and HARQ-F Model are originally introduced by Bollerslev et al. [BPQ16].

3.5 HAR-J

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \beta_J J_{t-1} + \mu_t$$

Another variation of the HAR model can be found at Andersen et al. [ABD07]. The HAR-J model is just like the standard HAR model, but with addition of an extra feature to account for discontinuous jumps in the data, $J_t \equiv \max(RV_t - BPV_t, 0)$. To calculate J_t, we need the Bi-Power Variation (BPV), which itself is a function of the daily returns, $BPV \equiv$

$\frac{\pi}{2}\sum_{i=1}^{M-1}|r_{t,i}||r_{t,i+1}|$. It is a measure that enables consistent estimations of continuous processes, even if there are jumps. This approach provides a decomposition into a continuous part – the genuine HAR model – and a discontinuous part J_t.

3.6 CHAR

$$RV_t = \beta_0 + \beta_1 BPV_{t-1} + \beta_2 BPV_{t-1|t-5} + \beta_3 BPV_{t-1|t-22} + \mu_t$$

The discontinuous part of the HAR-J model is known to be not very reliable though, so a different option to use the Bi-Power Variation in a model, would be to simply replace the RV-based components on the right-hand side with the BPV measure. This continuous HAR (CHAR) model is the only one among those that were tested, that does not retain the original structure of the simple AR(1) model in omitting all RV-based regressors. The model is still autoregressive though, as RV and BPV are both based on the intraday returns.

3.7 SHAR

$$RV_t = \beta_0 + \beta_1^+ RV_{t-1}^+ + \beta_1^- RV_{t-1}^- + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \mu_t$$

Another attempt to increase the performance of the HAR model was made by Patton and Sheppard [PS15], using the signed daily RV. Here, the RV is computed for positive and negative intraday returns separately, so that $RV_t^+ \equiv \sum_{i=1}^{M} r_{t,i}^2 1_{\{r_{t,i}^2>0\}}$ and $RV_t^- \equiv \sum_{i=1}^{M} r_{t,i}^2 1_{\{r_{t,i}^2<0\}}$. This model is called the Semivariance-HAR (SHAR) model.

4 Data

Symbol	*Min*	*Mean*	*Median*	*Max*	*AR*	*ARQ*
AXP	0.110	4.740	1.472	333.275	0.396	0.876
BA	0.107	2.693	1.450	234.095	0.236	0.727
CAT	0.186	3.425	1.787	121.328	0.754	1.013
CSCO	0.141	2.461	1.335	73.748	0.698	1.062
CVX	0.112	2.408	1.224	142.983	0.602	0.890
DD	0.094	2.947	1.410	369.797	0.162	0.687
DIS	0.121	3.053	1.143	1325.594	0.017	0.233
GE	0.116	3.123	1.123	177.557	0.683	1.072
HD	0.144	2.945	1.223	498.612	0.135	0.667
IBM	0.098	1.654	0.830	71.011	0.663	1.103
INTC	0.136	2.970	1.472	772.383	0.041	0.496
JNJ	0.078	1.368	0.591	335.130	0.036	0.317
JPM	0.136	5.092	1.684	253.366	0.671	1.122
KO	0.063	1.261	0.683	60.043	0.585	1.031
MCD	0.079	1.414	0.682	132.817	0.346	0.786
MMM	0.103	2.247	0.978	630.687	0.031	0.533
MRK	0.108	2.576	1.200	245.342	0.229	0.791
MSFT	0.151	2.102	1.228	61.790	0.693	1.025
NKE	0.141	2.666	1.367	129.502	0.361	0.732
PFE	0.145	2.196	1.143	567.817	0.039	0.567
PG	0.090	1.264	0.646	70.485	0.515	1.084
TRV	0.119	3.377	0.903	283.918	0.548	0.878
UNH	0.142	4.235	1.663	602.023	0.138	0.667
UTX	0.100	2.157	1.066	328.550	0.169	0.743
VZ	0.105	1.913	0.864	109.363	0.554	0.844
WMT	0.099	1.786	0.783	457.041	0.039	0.325
XOM	0.105	2.027	0.997	135.604	0.653	1.086

Table 1: Summary table for the Realized Variance (RV) of the 27 involved Dow Jones industrial average stocks. Columns AR and ARQ represent the β_1 coefficient of the respective models.

In Bollerslev et al. [BPQ16], the data used for comparing the performances of the forecasting models was the S&P 500 aggregate market index as well as the individual stocks of the Dow Jones industrial average with the exception of Apple, Goldman Sachs and Visa. With the sample starting on April 21, 1997 and ending on December 31, 2013 for the individual stocks of the Dow Jones industrial average, they had 3202 observations available for their analysis. Because of limited accessibility, we only used the 27 individual stocks of the DJIA and had to dispense with the S&P 500 Index. Also, our sample is considerably smaller: starting from June 27, 2007 and ending on June 26, 2017 we have exactly ten years of data available, what makes 2517 days of observation. Of these 2517 days, 13 showed incomplete or corrupt orderbook data and thus were excluded from the dataset, leaving 2504 days of observation to remain. Individual Stock Data was obtained from lobsterdata.com [Lob], providing several options of processing the data. The selected output contained 5-minute snapshots of level one order books for all stocks, what translates to 79 data points per day, where one data point carries information about the highest bid price and the lowest ask price at the time of a particular 5-minute snapshot of that day. Of the beginning 2517 days, 13 showed incomplete or corrupt orderbook data and thus were excluded from the dataset, leaving 2504 days of observation to remain. From the bid and ask prices, the logarithmic middle price was calculated as suggested in Dacoragna [DGM+01]. From there on, all calculation was performed as mentioned in Chapter 2. Table one contains descriptive statistics of the Realized Variance for all stocks across the complete sample. Additionally, the beta1 coefficients of the first order autoregressive model and the ARQ model have been calculated and added to the table. Most values are reasonable, however the maximum RV of the Walt Disney stock (DIS) of 1325.594 appears to be much larger than of all other stocks. This extreme value seems to be the result of a considerable increase of the ask price in the middle of the day on July 3, 2014. If we take a look at the beta1 coefficients of the AR and ARQ models, we can see that, for every stock, the ARQ coefficient is distinctly larger than the AR coefficient. This indicates that the coefficient of the RQ term in the ARQ model is indeed negative and balances out against the RV part.

5 Evaluation

To evaluate the models presented earlier, out of sample forecasts were conducted utilizing both, rolling window and increasing window. The window size of the rolling window was 1000 days and the increasing window also started with an interval of the same size to match the analysis of Bollerslev et al. [BPQ16]. Furthermore, two different loss-functions were used, namely the mean squared error and the QLIKE loss-function.

$$MSE(RV_t, X_t) \equiv (RV_t - X_t)^2 \qquad QLIKE(RV_t, F_t) \equiv \frac{RV_t}{F_t} - \log\left(\frac{RV_t}{F_t}\right) - 1$$

Obviously, the QLIKE loss-function is not capable of processing forecasting values with a different sign than the corresponding true value. Since the volatility is always non-negative, one can still have problems if the Realized Variance is zero on any specific day. However, some of the forecasting models did actually return negative forecasting values for a small number of days. Considering the negative coefficients for the RQ terms of the ARQ, HARQ and HARQ-F models and the rare extreme values for the RV as seen in Table 1, this is not very surprising. Interestingly, the HARJ, CHAR and SHAR models did produce a few negative forecasting values as well. As already mentioned, negative values are not possible for the Realized Variance, so one possibility to deal with this situation is to set those values to zero. This works perfectly fine for the MSE, but is not a solution for the QLIKE loss-function, since the logarithm of zero is not defined. This problem was not mentioned in Bollerslev et al. [BPQ16], so unfortunately it is not clear if they had to avoid negative forecasting values and how it was done, or if they didn't encounter any negative values at all. Yet, to still be able to attain QLIKE losses, the individual negative values were omitted specifically for this calculation. For the MSE, calculations were performed with negative values set to zero (Table 2), negative values kept as they were (Table 3) and with omitted negative values for better comparison to the QLIKE results (Table 4 and Table5).

In addition to the models that are to be compared in this analysis, four trivial estimators were also included in the process, which are three different weighted medians and a naïve estimator, which "predicts" the RV using the exact value of the RV of the last day. The weighted medians were calculated based on the subsamples that the rolling window and increasing window provided, so they used the same "training data" as the other models. Every ith weight was calculated by $w_i^{K,a} = a^i/(\sum_{l=0}^{K-1} a^l)$, where K equals the number of weights and parameter a determines the distribution of the weights and if set to 1, weights are such that the standard median is returned. For the evaluation, a was set to 0.95, 0.99 and 0.999.

Since the HAR Model appears to be the most popular model for volatility forecasting, all loss-values are standardized relative to the HAR model. In Table 2, where all negative forecasting values were set to zero, we already see very interesting behavior in the various models. While the HARQF model delivers the best performance for the stock average for the increasing window, we can see the reverse case for the rolling window. Since the median indicates that the HARQF model performs best for the rolling window as well, there must be one or more extreme outlier to produce an average of 3.0373 compared to the median of 0.9708. Of the trivial models, only the median with parameter $a = 0.999$ can overall outperform the HAR model, but does so exceedingly well, attaining the best average for the rolling window and the best median for the increasing window, the median outperforms every other model, except for HARQF in the rolling window median and the increasing window average.

MSE	*RW*		*IW*	
	Avg	*Med*	*Avg*	*Med*
AR	1.0138	1.0357	1.0907	1.0844
ARQ	1.0345	0.9757	0.9490	1.1286
HAR	1.0000	1.0000	1.0000	1.0000
HARQ	1.0766	0.9771	0.9542	1.0885
HARQF	3.0373	**0.9708**	**0.9450**	1.0423
HARJ	1.0027	0.9859	1.0000	0.9967
CHAR	0.9889	0.9815	0.9801	0.9926
SHAR	1.0006	1.0849	1.0099	1.2965
Med 0.95	1.0976	2.0637	0.9626	1.1513
Med 0.99	1.0195	1.0764	0.9578	1.0395
Med 0.999	**0.9785**	0.9765	0.9474	**0.9794**
naive	1.9344	1.8973	1.8713	1.9292

Table 2: Mean Squared Errors for rolling window and increasing window using average and median of the DJIA stocks. Negative forecasting values were set to zero.

Considering Table 3, where all results were calculated without changing negative forecasting values, it is instantly clear, that the ARQ, HARQ and HARQF models are the most impaired by not dealing with this problem. This needs to be addressed if using the Realized Quarticity in forecasting models, as these models are obviously prone to not only produce negative forecasting values but also to miss the true value by a large margin in the case of a negative value. Other models seem to either not produce any negative values, which of course is also the case for the trivial estimators, or to handle them in an acceptable manner.

MSE	*RW*		*IW*	
	Avg	*Med*	*Avg*	*Med*
AR	1.0138	1.0357	1.0907	1.0844
ARQ	9.3811	9.6375	14.0353	1.1286
HAR	1.0000	1.0000	1.0000	1.0000
HARQ	4.3724	5.2921	6.3465	1.0885
HARQF	7.3401	4.5635	5.8887	1.0423
HARJ	1.0045	0.9864	1.0112	1.1103
CHAR	0.9889	0.9815	0.9801	0.9926
SHAR	1.0012	1.0849	1.0102	1.2965
Med 0.95	1.0975	2.0637	0.9626	1.1513
Med 0.99	1.0195	1.0764	0.9578	1.0395
Med 0.999	**0.9784**	**0.9765**	**0.9474**	**0.9794**
naive	1.9343	1.8973	1.8713	1.9292

Table 3: Mean Squared Errors for rolling window and increasing window using average and median of the DJIA stocks. Negative forecasting values were treated as admissible and therefore left unchanged.

If the negative forecasting values are completely omitted (Table 4) as opposed to setting them to zero, the weighted median is still performing quite well, especially for the increasing window, but it is no longer the best choice in terms of MSE. Surprisingly enough, even omitting the worst cases for the ARQ, HARQ and HARQF models, they are still not consistently better than other models. There is, in fact, no model that can overall outperform the HAR model in all four categories. Comparing these results to the QLIKE losses, we can see that the HAR model is not contested by any other model (par HARQ for the rolling window median) in terms of this loss function. Considering these results, the popularity of the HAR model seems to be justified. The results don’t really match the numbers of the analysis in Bollerslev et al. [BPQ16], where the HARQ model broadly outperformed other models for the DJIA. To give more insight about the forecasting performances of the most interesting evaluated models, we look at the following Figures of a selection of stocks, that show forecast histograms and cumulative error plots.

MSE	*RW*		*IW*	
	Avg	*Med*	*Avg*	*Med*
AR	1.0010	1.0159	0.9976	1.0465
ARQ	1.0024	0.9959	0.9913	1.0631
HAR	1.0000	1.0000	1.0000	1.0000
HARQ	1.0219	0.9931	0.9927	1.0119
HARQF	1.1291	**0.9880**	**0.9882**	1.0018
HARJ	0.9996	0.9992	0.9965	1.0332
CHAR	**0.9974**	1.0006	0.9944	1.0215
SHAR	0.9996	1.0017	1.0018	**0.9970**
Med 0.95	1.1201	2.0863	1.0070	1.1632
Med 0.99	1.0420	1.0960	1.0021	1.1157
Med 0.999	1.0005	0.9963	0.9912	1.0353
naive	1.0476	1.0893	1.0263	1.1381

Table 4: Mean Squared Errors for rolling window and increasing window using average and median of the DJIA stocks. Negative forecasting values were omitted individually.

QLIKE	*RW*		*IW*	
	Avg	*Med*	*Avg*	*Med*
AR	1.0920	1.1813	1.2152	1.3745
ARQ	1.1651	1.0998	1.3385	1.2837
HAR	**1.0000**	1.0000	**1.0000**	**1.0000**
HARQ	1.0787	**0.9922**	1.1423	1.0945
HARQF	1.1571	1.4321	1.6134	1.8570
HARJ	1.3238	1.0068	1.0245	1.0666
CHAR	1.0954	1.0322	1.0708	1.0273
SHAR	1.0157	1.1028	1.0534	1.0740
Med 0.95	1.5316	1.8335	1.4909	1.7187
Med 0.99	1.4377	1.6028	1.3968	1.4958
Med 0.999	1.2631	1.3325	1.2111	1.3189
naive	1.9341	1.6374	2.0315	1.5344

Table 5: Mean QLIKE scores for rolling window and increasing window using average and median of the DJIA stocks. Negative forecasting values were omitted individually.

AXP: American Express

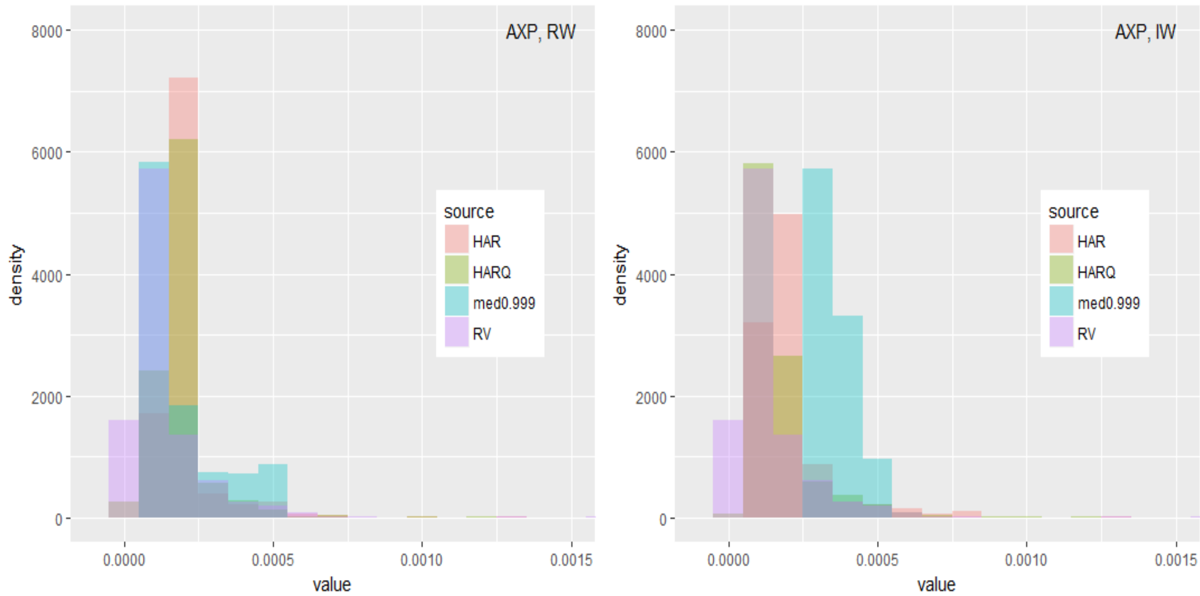

Figure 1: AXP histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

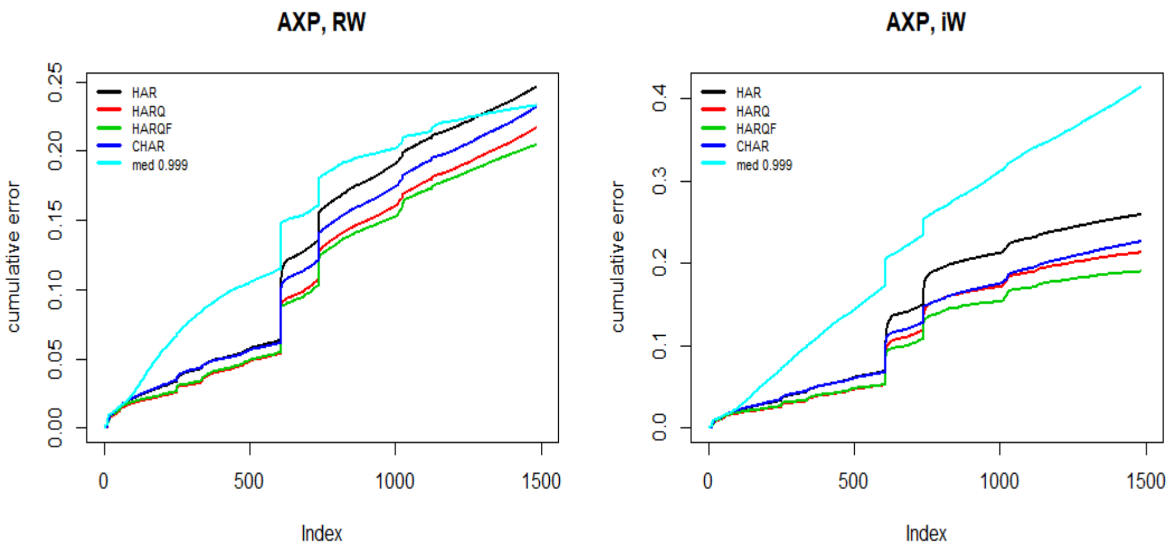

Figure 2: AXP cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

CAT: Caterpillar

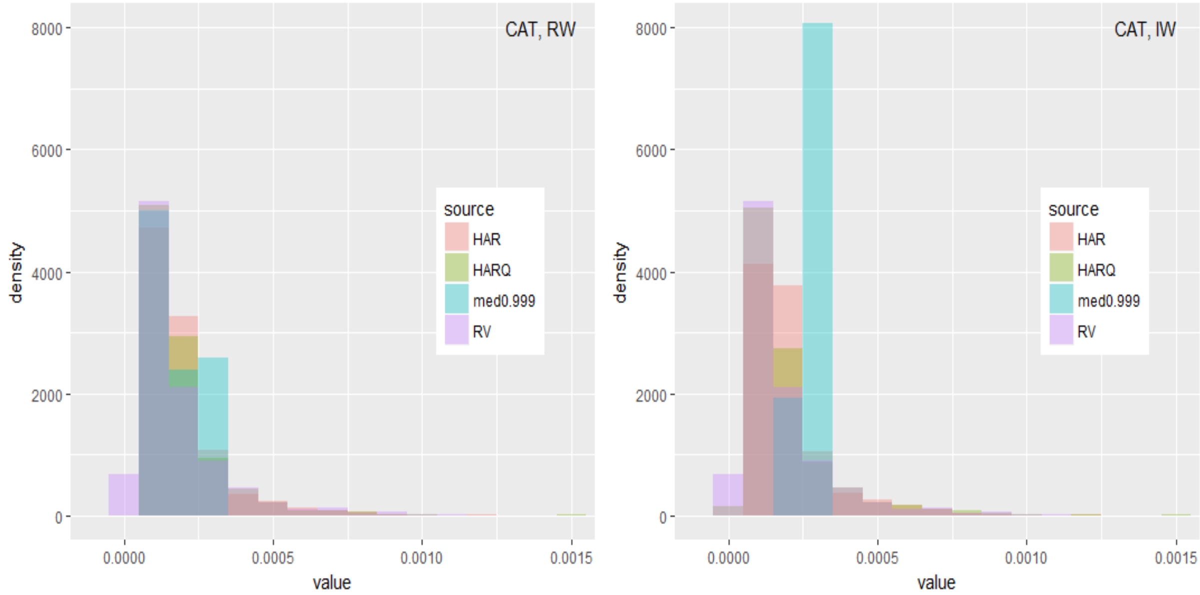

Figure 3: CAT histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

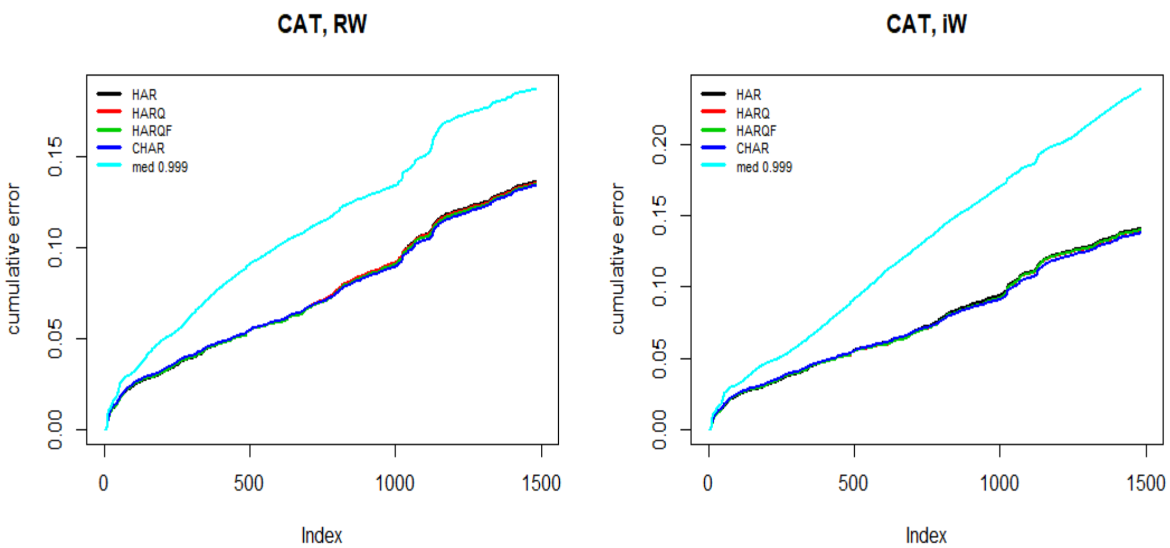

Figure 4: CAT cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

CSCO: Cisco Systems

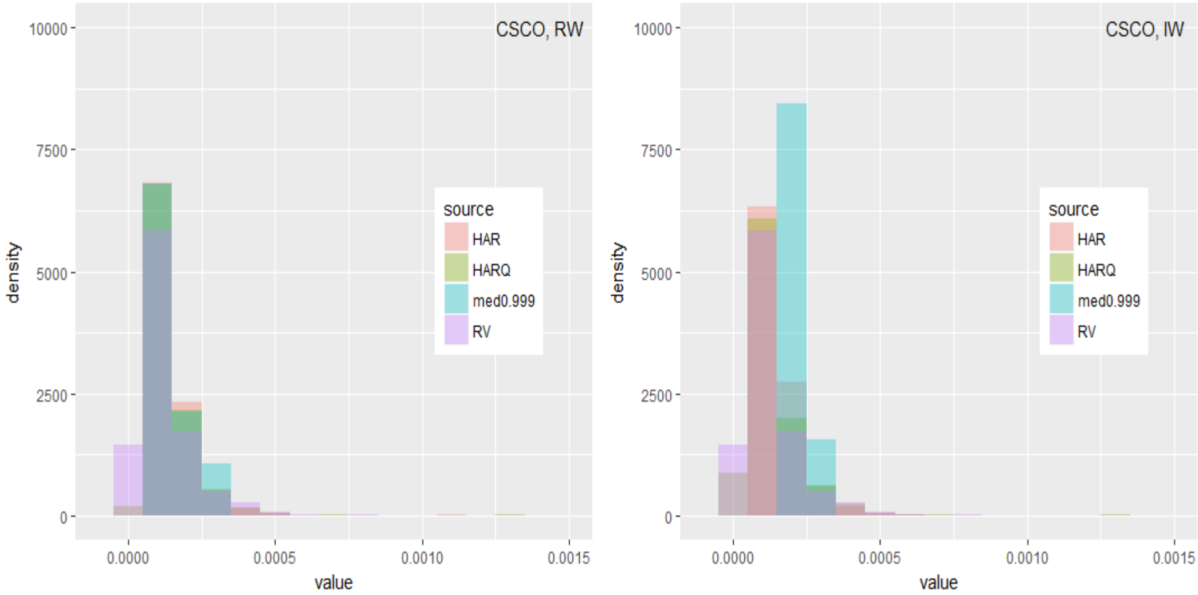

Figure 5: CSCO histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

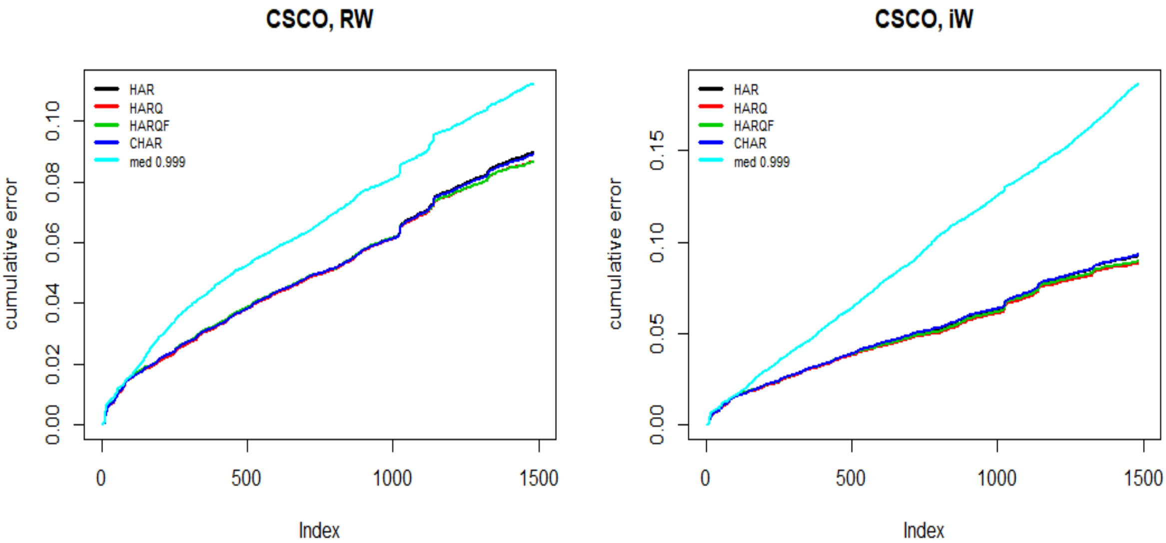

Figure 6: CSCO cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

DIS: Walt Disney

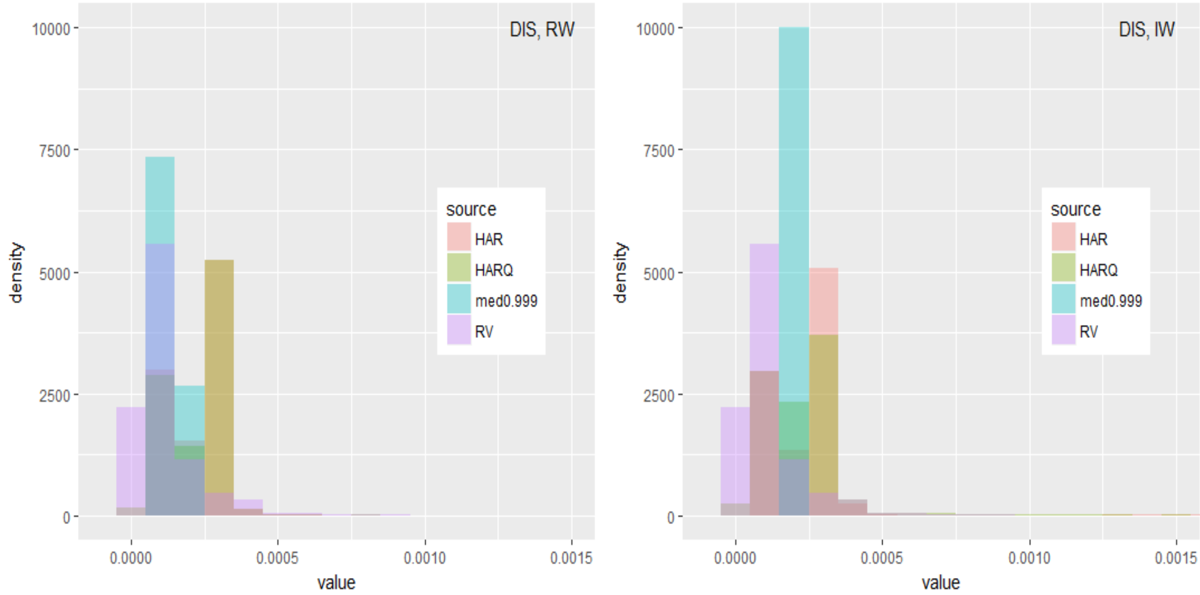

Figure 7: DIS histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

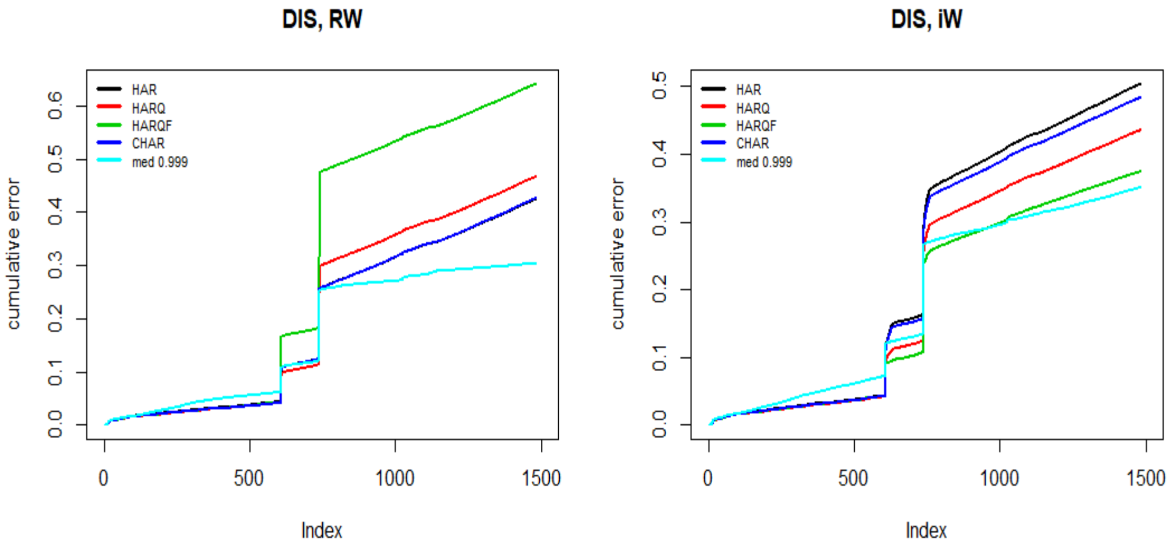

Figure 8: DIS cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

IBM: IBM

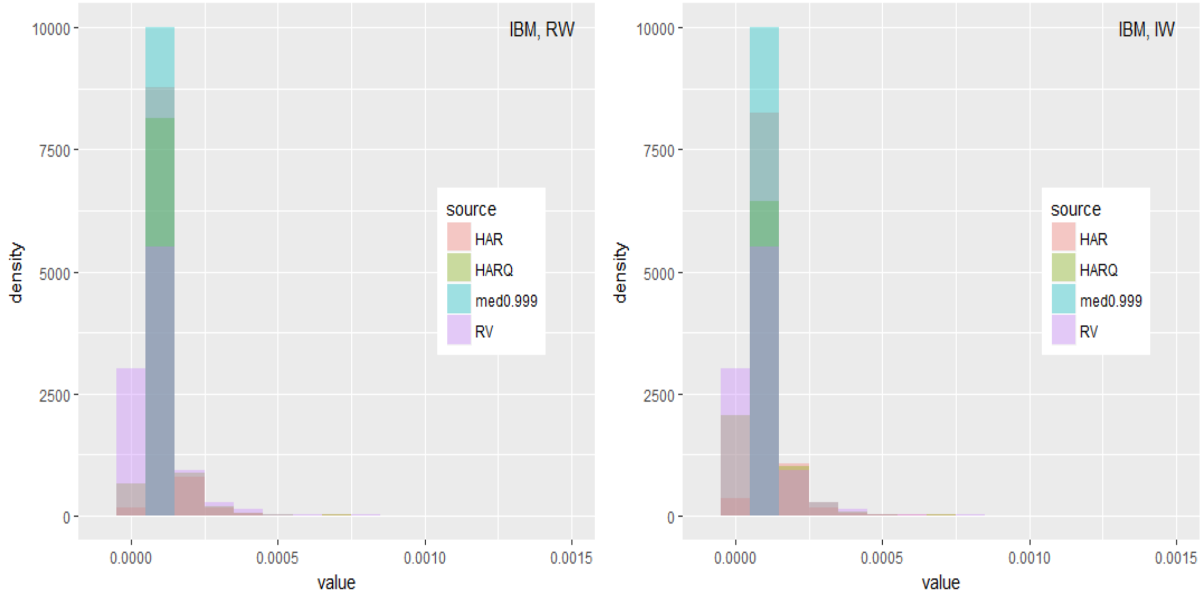

Figure 9: IBM histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

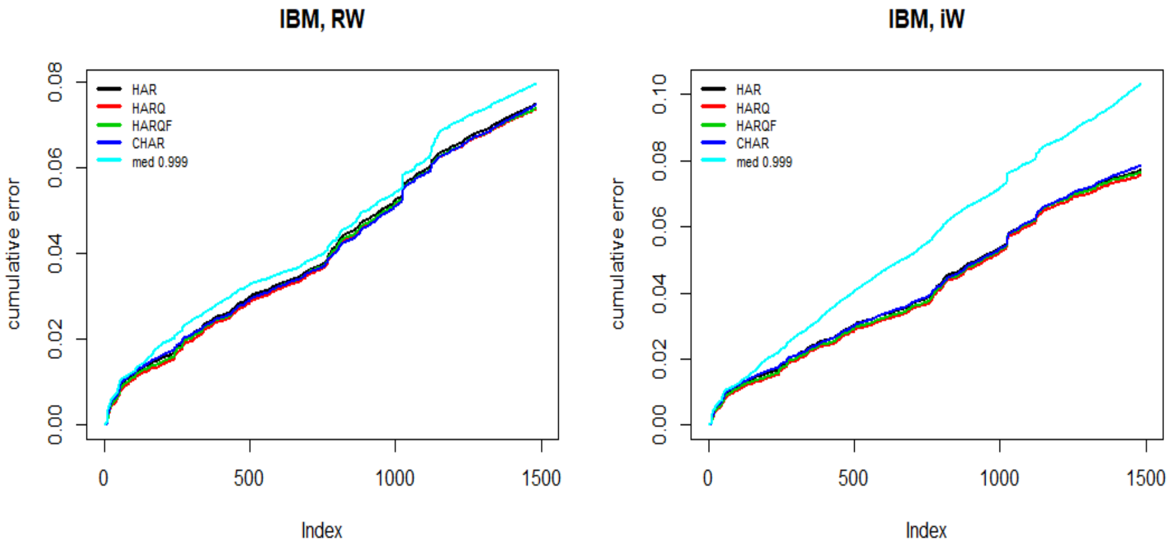

Figure 10: IBM cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

KO: Coca-Cola

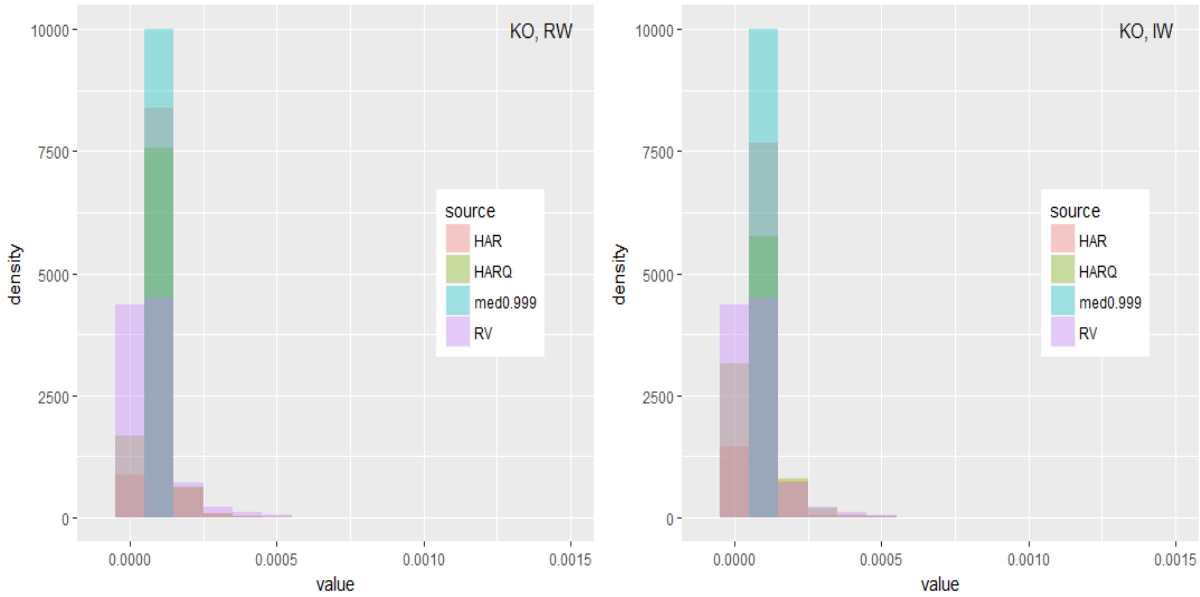

Figure 11: KO histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

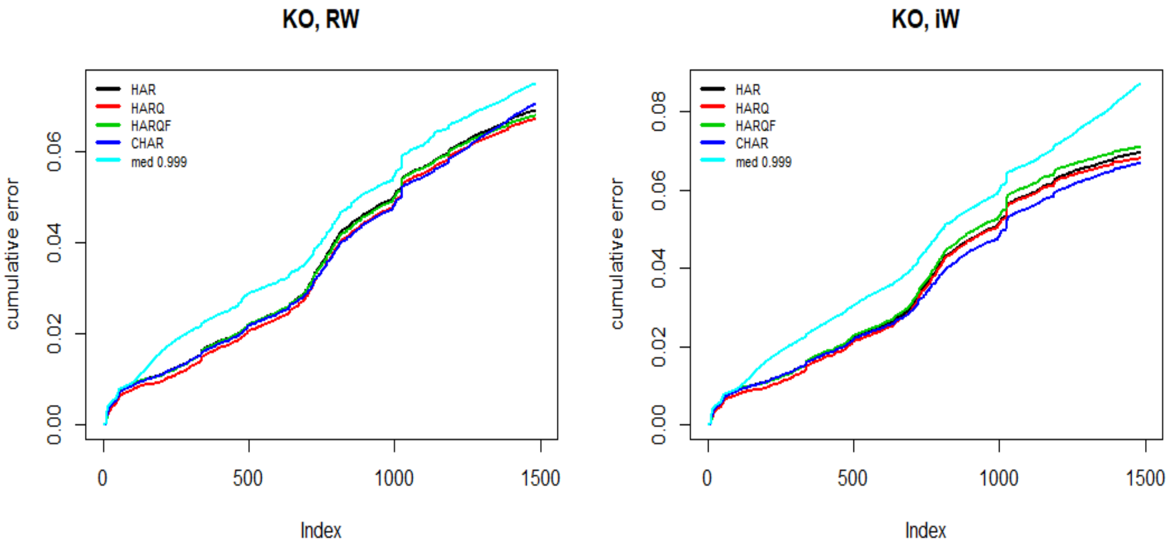

Figure 12: KO cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

MMM: 3M

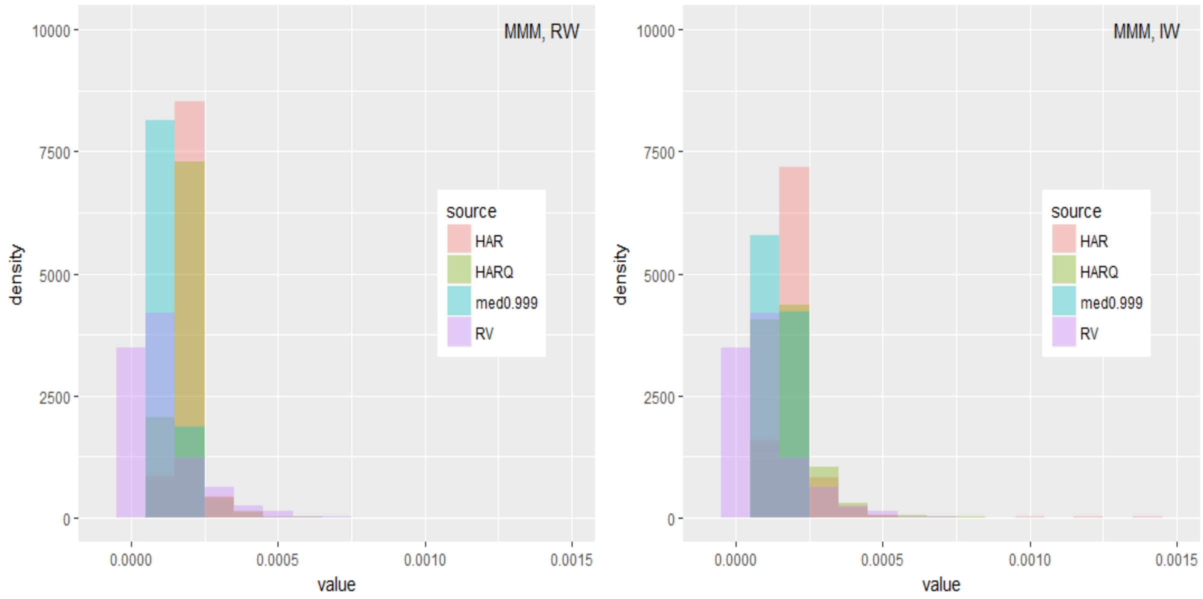

Figure 13: MMM histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

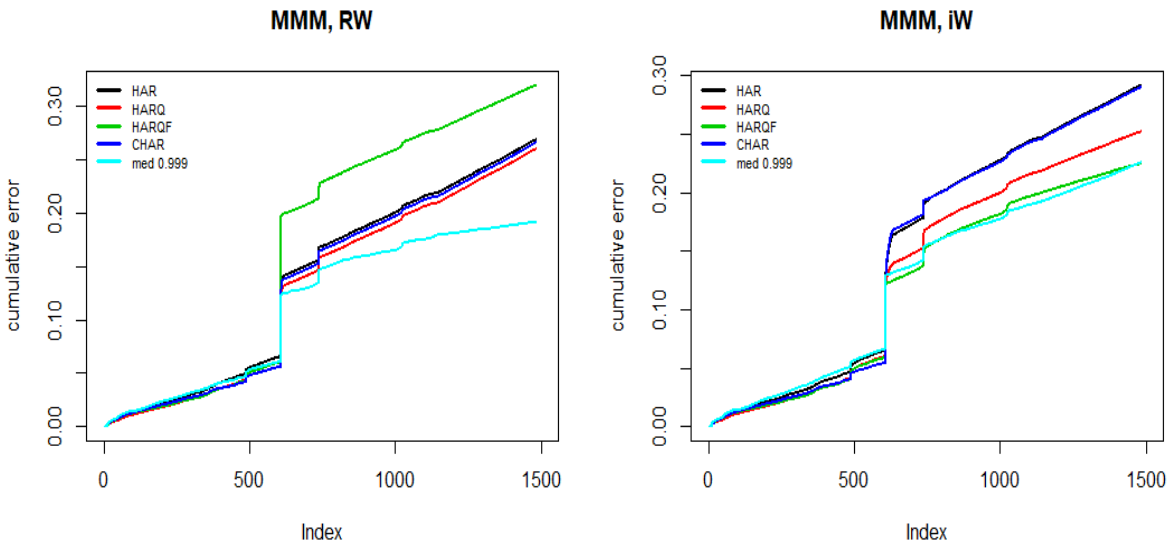

Figure 14: MMM cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

MSFT: Microsoft

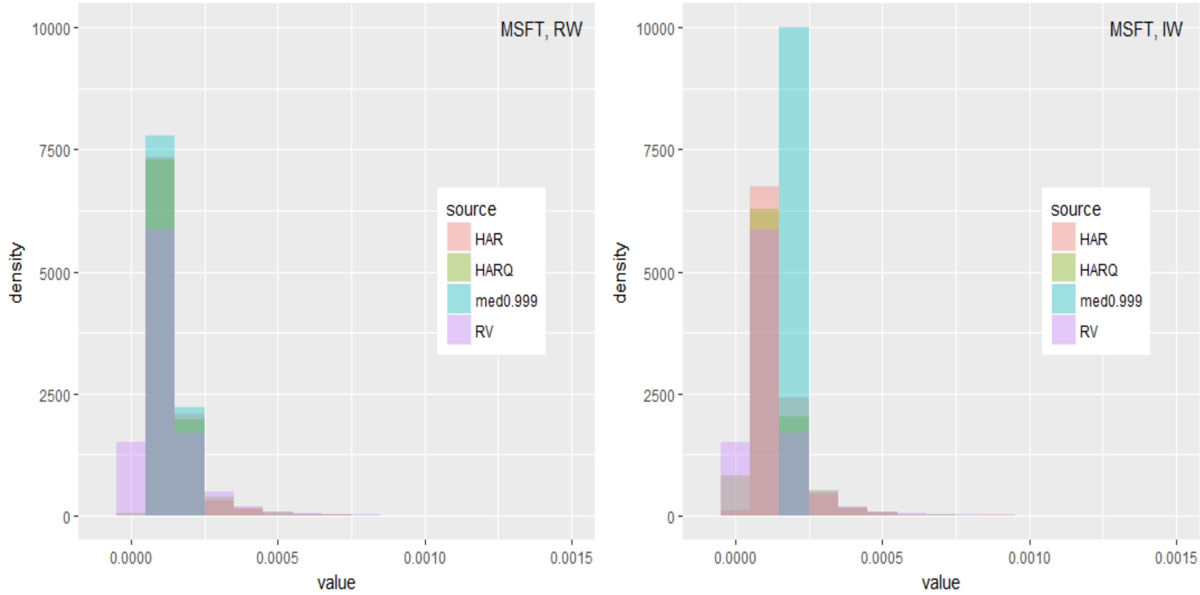

Figure 15: MSFT histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

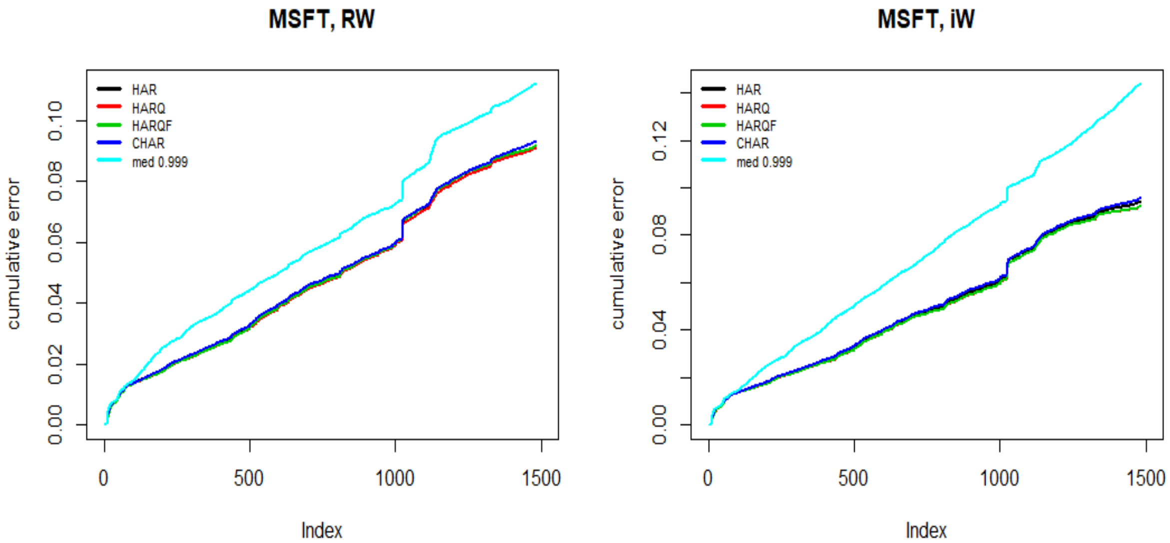

Figure 16: MSFT cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

PFE: Pfizer

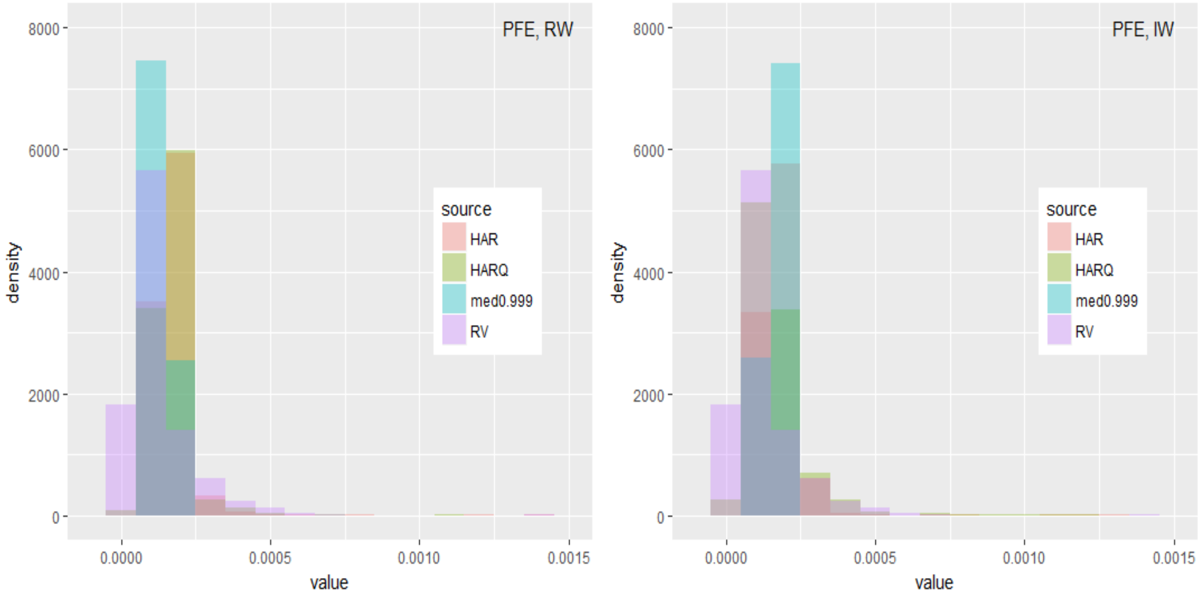

Figure 17: PFE histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

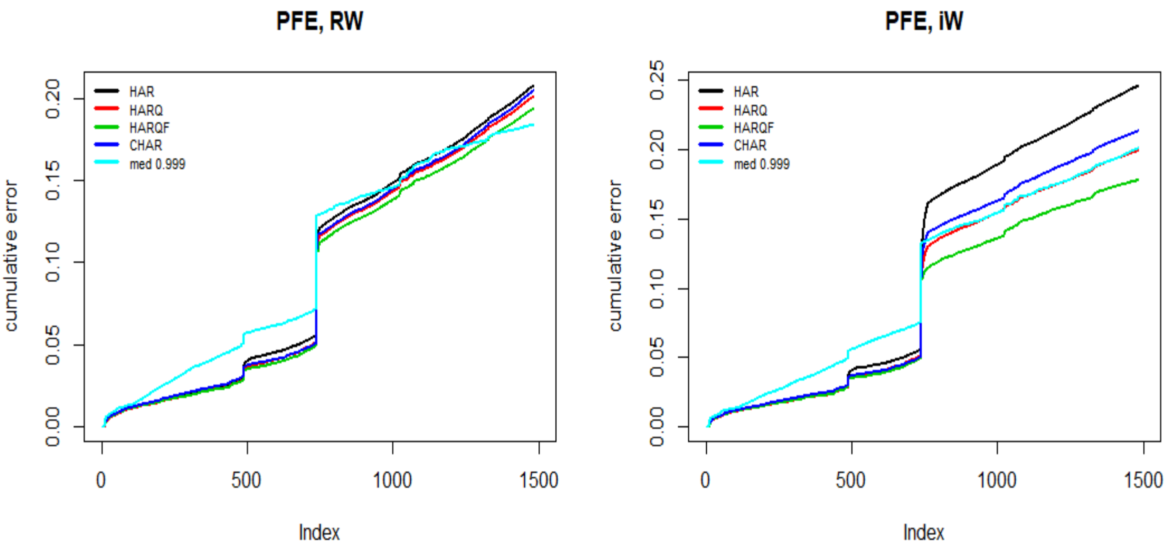

Figure 18: PFE cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

WMT: Wal-Mart

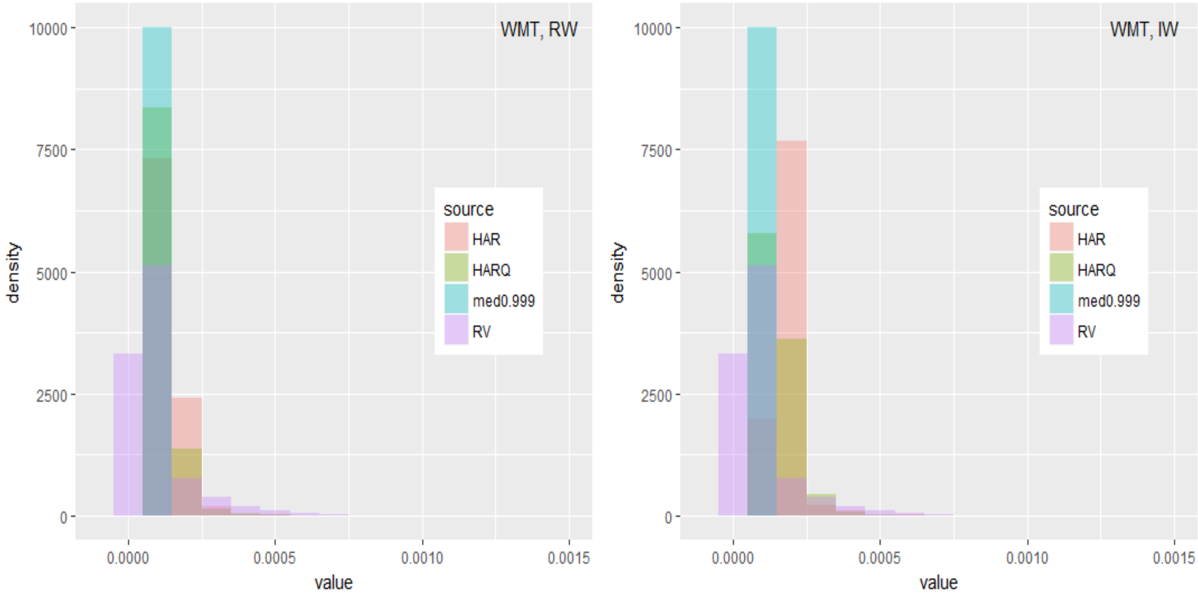

Figure 19: WMT histograms of RV and forecasts by HAR, HARQ and med 0.999 for RW and IW.

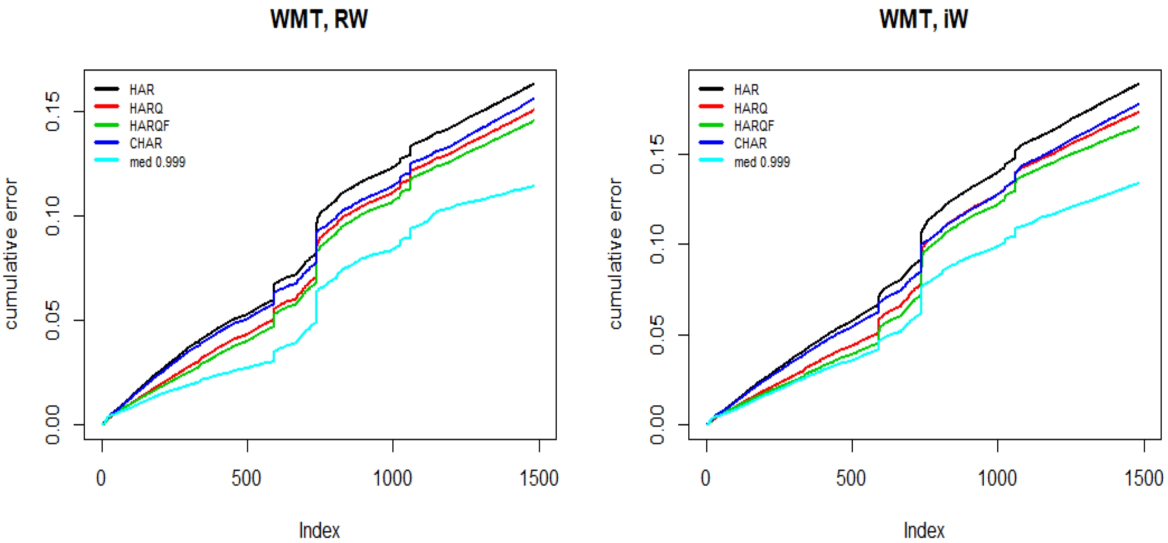

Figure 20: WMT cumulative absolute deviation of HAR, HARQ, HARQF, CHAR and med 0.999 for RW and IW.

The histograms show the density of the forecasting values of the HAR, HARQ and med 0.999 models, as well as the Realized Volatility, that was opted to be predicted by the models. Adding another forecasting model to the histograms leads to poor readability, so only the most interesting models were selected, including the HAR model for solid overall performance and popularity, the HARQ model as it performed best for Bollerslev et al. [BPQ16] and the median 0.999 since it is the best among the trivial estimators and performed very well in comparison to the other models. The cumulative error plots show the cumulative absolute deviation for the HAR, HARQ, HARQF, CHAR and med 0.999 models. The HARQF and CHAR model were added to these plots because the HARQF model already showed interesting behavior in the MSE tables and the CHAR model was the only model, that could consistently outperform the HAR model in the average MSE across all treatments of negative forecasting values. The left-hand plots show the results for the rolling window and the right-hand plots show the results for the increasing window. All plots were created with negative values set to zero.

The first thing that is immediately observable, is that the models behave very differently across the selected stocks. Especially the weighted median and the HARQF model are in many cases performing either distinctly better or worse than the other models (compare for example the cumulative error plots of DIS and MMM). The other three models perform very similarly, with the CHAR and HARQ models performing overall better than the HAR model. While the HARQF model appears to be able to make good use of the additional information that the increasing window provides, the weighted medians performance is impeded by this method. The HAR, HARQ and CHAR models show almost identical results for the rolling window and the increasing window. The findings derived from the cumulative error plots do correspond with the histograms, considering the models that are included in both graphics. One issue that the HAR, HARQ and median 0.999 have in common, is that they systematically overestimate the Realized Variance, indicated by the leftmost purple bar at the zero mark on the x-axis. The weighted median tends to heavily overestimate the Realized Variance, especially for the increasing window. To a lesser extent, this is also the case for the HAR model. The HARQ forecasts seem mostly inconspicuous in comparison, which is also the case for the cumulative errors of the HARQ model.

6 Conclusion

A detailed graphical analysis was produced to better understand the differences of the several models that were evaluated in this study. For reasons of simplification and readability, the focus has been put on the HAR, HARQ, HARQF, CHAR and med 0.999 models. While the HARQF and med 0.999 models scored best in terms of MSE - depending on how negative values were treated – they both showed unsettling inconsistent performance across the individual stocks. The HAR, HARQ and CHAR models appeared to be more robust and delivered similar results for all stocks and also the rolling window and increasing window alike.

However, the significant success of the HARQ model, that Bollerslev et al. [BPQ16] found in their analysis, could not be reproduced. If anything, one could say that the HARQ model is maybe on par with the HAR model, which scored best in the QLIKE loss function, even though the negative forecasting values were omitted for this calculation. This is of course another issue if dealing with the “Q” models: they are likely to produce negative forecasting values which must be taken care of.

Finally, the CHAR model appears as very stable and performing quite well across all stocks, for rolling window and increasing window alike. It is very close to the ARQ model, and can do so without including another term for the Realized Quarticity. Thus, the CHAR model would be the recommended method for forecasting the volatility of the Dow Jones industrial average. The behavior of the Realized Quarticity in the context of volatility forecast is left to be subject to further research.

7 Appendix

The following functions were created to calculated the variables required by the forecasting models. These are the logarithmic middle price, the intraday returns, the daily Realized Variance, the daily Realized Quarticity, the daily bi-power variation, the daily positively signed Realized Variance, the daily negatively signed Realized Variance and the daily jump variation.

```
log.price <- function(x, AP=1, BP=3) {
  if (nrow(x)>0){
    x        <-  x[!rowSums(x<0),]
  }
  p    <-  (log(x[,AP])+log(x[,BP]))/2
  p
}


intra.return <- function(p) {
  r        <-  p[-1] - p[-length(p)]
  r
}


daily.rv <- function(r) {
  rv       <-  sum(r^2)
  rv
}


daily.rq <- function(r) {
  rq       <-  (length(r)/3)*sum(r^4)
  rq
}


daily.bpv <- function(r) {
  bpv      <-  (pi/2)*sum(abs(r[-1])*abs(r[-length(r)]))
  bpv
}


daily.rv.pos <- function(r) {
  r           <-  r[(r>0)]
  rv          <-  sum(r^2)
  rv
}


daily.rv.neg <- function(r) {
  r           <-  r[(r<0)]
  rv          <-  sum(r^2)
  rv
}


daily.j <- function(RV, BPV){
  j      <- sapply(RV-BPV, function(H){ max(H,0) })
  j
}
```

The above functions were then implemented in a function to calculate a data frame containing all variables required by the forecasting models including the weekly and monthly Realized Variance, Realized Quarticity and bi-power variation, as well as the RV and RQ cross terms for the “Q” models. Input for this function is a list of Orderbooks as they are acquired from lobsterdata.com [Lob].

```
calc.data <- function(OB){
    p     <- lapply(OB, log.price)
    r     <- lapply(p, intra.return)
    RV    <- sapply(r, daily.rv)
    RQ    <- sapply(r, daily.rq)
    BPV   <- sapply(r, daily.bpv)
    J     <- daily.j(RV, BPV)
    RVpos <- sapply(r, daily.rv.pos)
    RVneg <- sapply(r, daily.rv.neg)

    #(t)
    pRV    <-    RV[-c(1:21, 22)]

    #(t-1)
    nd <-length(RV)
    dRV    <-    RV[-c(1:21, nd)]
    dRQ    <-    RQ[-c(1:21, nd)]
    dBPV   <-   BPV[-c(1:21, nd)]
    dJ     <-     J[-c(1:21, nd)]
    dRVpos <- RVpos[-c(1:21, nd)]
    dRVneg <- RVneg[-c(1:21, nd)]
    dRVRQ  <- dRV*sqrt(dRQ)

    #(t-1|t-5)
    nw <- (nd-4)
    wRV    <- rollmean(RV , k=5 )[-c(1:17, nw)]
    wRQ    <- rollmean(RQ , k=5 )[-c(1:17, nw)]
    wBPV   <- rollmean(BPV, k=5 )[-c(1:17, nw)]
    wRVRQ  <- wRV*sqrt(wRQ)

    #(t-1|t-22)
    nm <- (nd-21)
    mRV    <- rollmean(RV , k=22)[-nm]
    mRQ    <- rollmean(RQ , k=22)[-nm]
    mBPV   <- rollmean(BPV, k=22)[-nm]
    mRVRQ  <- mRV*sqrt(mRQ)

    #compose the calculated and prepared data to a dataframe
    DATA <- cbind.data.frame(pRV, dRV, dRQ, dBPV, dJ, dRVpos, dRVneg,
                  wRV, wRQ, wBPV, mRV, mRQ, mBPV, dRVRQ, wRVRQ, mRVRQ)
    DATA
}
```

The data frame created by the function above then serves as the input data for the following two functions, which create forecasting values for several models based on a rolling window and increasing window respectively. The parameter k denotes the size of the rolling window, and the starting size of the increasing window, respectively.

```
rw.fc <- function(DATA, k=1000){
    j <- 0:(k-1)
    a95  <-  0.95^j ;  w95 <- a95/sum(a95) #create weights for weighted median 0.95
    a99  <-  0.99^j ;  w99 <- a99/sum(a99) #create weights for weighted median 0.99
    a999 <- 0.999^j ; w999 <- a999/sum(a999) #create weights for weighted median 0.999

    reps  <- dim(DATA)[1]-k
    preds <- matrix(0, nrow = reps, ncol = 12)
    colnames(preds) <- c("AR", "ARQ", "HAR", "HARQ", "HARQF", "HARJ", "CHAR", "SHAR",
         "med_0.95", "med_0.99", "med_0.999", "naive")
    for (i in 1:reps){
        AR.fit    <- lm(pRV ~ dRV                                                  ,
data = DATA, subset = i:(k-1+i))
        ARQ.fit   <- lm(pRV ~ dRV + dRVRQ                                          ,
data = DATA, subset = i:(k-1+i))
        HAR.fit   <- lm(pRV ~ dRV              + wRV           + mRV               ,
data = DATA, subset = i:(k-1+i))
        HARQ.fit  <- lm(pRV ~ dRV + dRVRQ      + wRV           + mRV               ,
data = DATA, subset = i:(k-1+i))
        HARQF.fit <- lm(pRV ~ dRV + dRVRQ      + wRV + wRVRQ   + mRV + dRVRQ       ,
data = DATA, subset = i:(k-1+i))
        HARJ.fit  <- lm(pRV ~ dRV              + wRV           + mRV           + dJ ,
data = DATA, subset = i:(k-1+i))
        CHAR.fit  <- lm(pRV ~ dBPV             + wBPV          + mBPV              ,
data = DATA, subset = i:(k-1+i))
        SHAR.fit  <- lm(pRV ~ dRVpos + dRVneg  + wRV           + mRV               ,
data = DATA, subset = i:(k-1+i))

        AR.pred    <- predict(AR.fit,    newdata = DATA[k+i,])
        ARQ.pred   <- predict(ARQ.fit,   newdata = DATA[k+i,])
        HAR.pred   <- predict(HAR.fit,   newdata = DATA[k+i,])
        HARQ.pred  <- predict(HARQ.fit,  newdata = DATA[k+i,])
        HARQF.pred <- predict(HARQF.fit, newdata = DATA[k+i,])
        HARJ.pred  <- predict(HARJ.fit,  newdata = DATA[k+i,])
        CHAR.pred  <- predict(CHAR.fit,  newdata = DATA[k+i,])
        SHAR.pred  <- predict(SHAR.fit,  newdata = DATA[k+i,])
        med_0.95 <- weightedMedian(x = DATA$pRV[i:(k-1+i)], w = w95)
        med_0.99 <- weightedMedian(x = DATA$pRV[i:(k-1+i)], w = w99)
        med_0.999 <- weightedMedian(x = DATA$pRV[i:(k-1+i)], w = w999)
        naive.pred <- DATA$pRV[k-1+i]

        preds[i,] <- c(AR.pred, ARQ.pred, HAR.pred, HARQ.pred, HARQF.pred, HARJ.pred,
         CHAR.pred, SHAR.pred, med_0.95, med_0.99, med_0.999, naive.pred)
    }
    data.frame(preds)
}
```

```
iw.fc <- function(DATA, k=1000){
    reps  <- dim(DATA)[1]-k
    preds <- matrix(0, nrow = reps, ncol = 12)
    colnames(preds) <- c("AR", "ARQ", "HAR", "HARQ", "HARQF", "HARJ", "CHAR", "SHAR",
        "med_0.95", "med_0.99", "med_0.999", "naive")
    for (i in 1:reps){
        AR.fit    <- lm(pRV ~ dRV                                                      ,
data = DATA, subset = 1:(k-1+i))
        ARQ.fit   <- lm(pRV ~ dRV + dRVRQ                                              ,
data = DATA, subset = 1:(k-1+i))
        HAR.fit   <- lm(pRV ~ dRV              + wRV           + mRV                   ,
data = DATA, subset = 1:(k-1+i))
        HARQ.fit  <- lm(pRV ~ dRV + dRVRQ      + wRV           + mRV                   ,
data = DATA, subset = 1:(k-1+i))
        HARQF.fit <- lm(pRV ~ dRV + dRVRQ      + wRV + wRVRQ   + mRV + dRVRQ           ,
data = DATA, subset = 1:(k-1+i))
        HARJ.fit  <- lm(pRV ~ dRV              + wRV           + mRV              + dJ ,
data = DATA, subset = 1:(k-1+i))
        CHAR.fit  <- lm(pRV ~ dBPV             + wBPV          + mBPV                  ,
data = DATA, subset = 1:(k-1+i))
        SHAR.fit  <- lm(pRV ~ dRVpos + dRVneg  + wRV           + mRV                   ,
data = DATA, subset = 1:(k-1+i))

        AR.pred    <- predict(AR.fit,    newdata = DATA[k+i,])
        ARQ.pred   <- predict(ARQ.fit,   newdata = DATA[k+i,])
        HAR.pred   <- predict(HAR.fit,   newdata = DATA[k+i,])
        HARQ.pred  <- predict(HARQ.fit,  newdata = DATA[k+i,])
        HARQF.pred <- predict(HARQF.fit, newdata = DATA[k+i,])
        HARJ.pred  <- predict(HARJ.fit,  newdata = DATA[k+i,])
        CHAR.pred  <- predict(CHAR.fit,  newdata = DATA[k+i,])
        SHAR.pred  <- predict(SHAR.fit,  newdata = DATA[k+i,])
        j <- 0:(k-1+(i-1))
        a95 <- 0.95^j ; w95 <- a95/sum(a95) #create weights for weighted median 0.95
        a99 <- 0.99^j ; w99 <- a99/sum(a99) #create weights for weighted median 0.99
        a999 <- 0.999^j ; w999 <- a999/sum(a999) #create weights for weighted median 0.999
        med_0.95 <- weightedMedian(x = DATA$pRV[1:(k-1+i)], w = w95)
        med_0.99 <- weightedMedian(x = DATA$pRV[1:(k-1+i)], w = w99)
        med_0.999 <- weightedMedian(x = DATA$pRV[1:(k-1+i)], w = w999)
        naive.pred <- DATA$pRV[k-1+i]

        preds[i,] <- c(AR.pred, ARQ.pred, HAR.pred, HARQ.pred, HARQF.pred, HARJ.pred,
         CHAR.pred, SHAR.pred, med_0.95, med_0.99, med_0.999, naive.pred)
    }
    data.frame(preds)
}
```

8 References

[ABD07] T. G. Andersen, T. Bollerselv and F. X. Diebold. Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. Review of Economics and Statistics, 89(4): 701-720, 2007.

[BPQ16] T. Bollerslev, A. J. Patton and R. Quaedvlieg. Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics, 192(1): 1-18, 2016.

[Cor09] F. Corsi. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7(2): 147-196, 2009.

[DGM+01] M. M. Dacoragna, R. Gencay, U. A. Müller, R. B. Olsen and O. V. Pictet. An Introduction to High-Frequency Finance. 2001. Academic Press.

[Der11] E. Derman. Models.Behaving.Badly.: Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life. 2011. Free Press.

[GT07] D. G. Goldstein and N. N. Taleb. We don't quite know what we are talking about when we talk about volatility. Journal of Portfolio Management, 33(4): 84-86, 2007.

[Lob] www.lobsterdata.com

[PS15] A. J. Patton and K. Sheppard. Good Volatility, Bad Volatility: Signed Jumps and The Persistence of Volatility. Review of Economics and Statistics, 97(3): 683-697, 2015.

9 Abstract

Abstract (English):

When it comes to volatility forecasting, there are plenty models that come into question. One of the most recent is the so called HARQ model, by Bollerslev et al. [BPQ16]. It considers the Realized Quarticity, trying to compensate large forecasting errors caused by a volatile stock market. This is a new approach in the field of volatility forecasting and therefore an optimal subject for a replication study.

An extended evaluation was conducted, reproducing the analysis of Bollerslev et al. [BPQ16] as well as a graphical analysis to complement the results. To forecast the Realized Variance of the Dow Jones industrial average, the models were trained by both, rolling windows and increasing windows. The models were then compared using the MSE and QLIKE loss-functions. For the graphical analysis, histograms of the forecasting values and cumulative error plots were created.

While the HARQ model did do well overall, it did not perform distinctively better than more popular models such as the HAR and CHAR models. The success of the HARQ model could not be reproduced and the behavior of the Realized Quarticity in the context of volatility forecasting is left to be subject to further research.

Zusammenfassung (Deutsch):

In der Disziplin des volatility forecasting kommen eine Reihe von Modellen in Frage. Eines der aktuellsten ist das HARQ Modell von Bollerselv et al. [BPQ16]. Es berücksichtigt die Realized Quarticity, um größere Prognosefehler durch hohe Volatilität auf dem Aktienmarkt zu kompensieren. Dies ist ein neues Konzept im Bereich des volatility forecasting und daher bestens geeignet für eine Wiederholungsstudie.

Es wurde eine erweiterte Evaluierung durchgeführt, um die Analyse von Bollerslev et al. [BPQ16] zu reproduzieren, wie auch eine graphische Analyse um die Ergebnisse zu komplementieren. Um die Realized Variance des Dow Jones industrial average vorherzusagen, wurde training-data durch rolling window und increasing window erzeugt. Anschließend wurden die Modelle mittels der MSE und QLIKE Verlustfunktionen verglichen. Für die graphische Analyse wurden Histogramme der Vorhersagewerte und kumulierte Fehler plots erstellt.

Das HARQ Modell konnte zwar gut abschneiden, war im Vergleich allerdings nicht maßgeblich besser als andere, etablierte Modelle wie unter anderem das HAR oder CHAR Modell. Der Erfolg des HARQ Modells konnte somit nicht reproduziert werden und das Verhalten der Realized Quarticity im Kontext von volatility forecasting bleibt Gegenstand für weitere Untersuchungen.