DAILY VOLUME FORECASTING USING HIGH FREQUENCY PREDICTORS

Leandro G. M. Alvim
Departamento de Informatica
Pontificia Universidade Catolica do Rio de Janeiro
Rio de Janeiro, RJ, Brazil
email: leandrouff@gmail.com

Cicero N. dos Santos
Departamento de Informatica
Pontificia Universidade Catolica do rio de Janeiro
Rio de Janeiro, RJ, Brazil
email:

Ruy L. Milidiu
Departamento de Informatica
Pontificia Universidade Catolica do rio de Janeiro
Rio de Janeiro, RJ, Brazil
email: milidiu@inf.puc-rio.br

ABSTRACT
Daily volume is an important feature when it comes to financial market structure. Effective daily volume forecasting can help areas such as portfolio management and algorithm trading. Intraday updates of daily volume forecasts can explore high frequency data to provide more accurate forecasts. Previous work on daily volume forecasting usually use Bayesian methods. In our work, we approach the problem of daily volume forecasting using the intraday information. Forecasting is accomplished by the use of two machine learning predictors: Support Vector Regression (SVR) and Partial Least Squares (PLS). We empirically test our method using the top nine high liquidity Bovespa traded stocks. Our metrics are the percentage error and the relative error reduction against a naive strategy. Our results show that SVR and PLS provide accurate forecasts. Moreover, the forecasting accuracy improves throughout the day as more intraday information is available.

KEY WORDS
Finance, Volume Forecasting, Machine Learning, PLS, SVR.

1 Introduction

Investors who want to minimize their execution order market impact have extensively investigated the execution process and its consequences in exchanging amounts of assets. With the rising of the algorithm trading area, a lot of institutional investors are using computer based algorithms and pattern recognition techniques. It is quite rare to find human intervention in order execution. Nowadays, the total amount of orders executed by computer-based traders is large and increasing. According to cf. Chordia et al. (2008), algorithm trading has been reducing the average amount of trades in the market and institutional investors are forced to split their orders for better execution prices. One strategy is the Volume Weighted Average Price (VWAP) strategy. Its intention is to splits an specific number of shares into smaller number of orders during the day, executing them at different prices. Such splitting procedure aims at operating close to the VWAP price. One of the interesting aspects of this strategy is that accurate volume forecasting can lead to accurate VWAP execution.

Portfolio management and asset allocation require acquisition or liquidation of positions, placing large amounts of orders that could change the price of an asset. This change is strictly associated with the transaction risk and can result in lower profits or higher losses. There is no simple solution for this problem and to minimize this, the investor can take into account: asset volume, financial market rules, volatility and asset correlations for example.

Besides the importance of volume forecasting to algorithm trading and portfolio management, few works about volume can be found in the finance literature. In [Biakowski et al., 2005] and [Bialkowski et al., 2006] a new methodology is proposed. It consists of decomposing volume for intraday volume forecasting. Their work used ARIMA and SETAR models, which allowed significant reduction in vwap orders risk. Their data consisted in forty stocks of the CAC40 index. In [Lean et al., 2008] a new kernel-based ensemble learning approach is proposed. They use econometric models and artificial intelligence so as to predict China foreign trade volume. In [Lux and Kaizoji, 2007] the predictability of Japanese daily volume stocks and volatility are investigated. They compare ARFIMA and FIGARCH long-memory models to GARCH and ARIMA short-memory ones in order to predict the volume of 100 days ahead.

Our main contribuition is a dynamic volume model that uses high-frequency machine learning predictors for daily volume forecasting. In our work, we have updated our model aiming at reducing the daily forecasting error during the day. The updating mechanism model works using the already known intraday volume during the day. In order to forecast the volume during the day, we use a Support Vector Regression (SVR) and Partial Least Squares (PLS).

For the experiments we use the Bovespa data set. This data set contains $15-$min intraday volume of 9 stocks in which three of them consist of high liquidity stocks and the remaining are low liquidity stocks. Table 1 shows the average percentage error reduction of our proposed method against the naive predictor, with a intraday volume model. The collumns of table 1, represent the performance gain in an interval of $15min$. The collumns represent the average performance gains for the beginning, middle and final of the market day. For all these three periods, we use the prior intraday information of the model that uses intraday information. In table 1, our proposed method outperforms the traditional one, with and without prior intraday information during the day. These results lead us to conclude that the use of machine learning techniques have improved a volume predictor performance since the beginning of the day.

Stock	intervals 1-9	intervals 10-18	intervals 19-27
PETR3	0,28	0,36	0,51
PETR4	0,22	0,38	0,46
VALE5	0,24	0,37	0,45
CSNA3	0,55	0,44	0,45
USIM5	0,19	0,20	0,42
BBDC4	0,27	0,41	0,44
ITSA4	0,34	0,34	0,52
ITAU4	0,26	0,30	0,44
GOAU4	0,23	0,26	0,42

Table 1. Average relative error reduction against a naive strategy

Our results are useful for investors who want to choose the convenient time during the day to execute orders according to a threshold error. By using a VWAP strategy, investors are able to minimize the market impact of their order execution. Moreover, our model can help other intraday and interday forecasting models such as price, volatility and intraday volume.

This work is organized in the following way: in section 2, we present the intraday and interday volume dynamics and propose our dynamic volume model; in section 3, we describe the PLS and SVR machine learning predictors; in section 4, we show our empirical methodology used in the experiments; in section 5, we compare our results with a naive strategy; and, in section 6, our conclusions and future works are presented.

2 Modeling Volume Dynamics

Despite this work deals with daily volume forecasting, we use an intraday volume information to support daily predictions. Taking into account this intraday information, we now investigate the peculiarities of intraday behaviour.

It is common knowledge that seasonal fluctuations occurs on intraday volume. In [Biais et al., 1995] they show that exist an U-shape seasonal pattern. To circumvent this problem [Dufour and Engle, 2000] and [Gouriroux and Fol, 1998] worked on market time scale instead of calendar time scale. Another approach, proposed by [Engle and Russel, 1998] and [Easley and OHara, 1987], uses a volume correction on a stock-by-stock time varying average. In [McCulloch, 2004] the time varying and an average volume across stocks are used. We can see in our data set, in figure 1, an example of the $U-$Shape pattern.

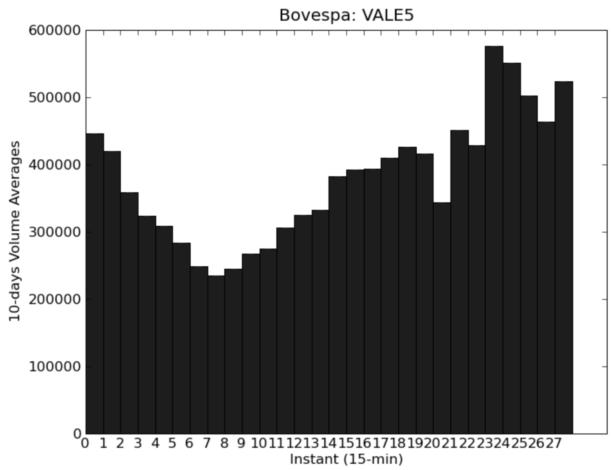

Figure 1. VALE5 U-shape pattern.

Taking into account the seasonality problem and the $U-$Shape pattern of the intraday volume information, we construct specialized models for each of the market day interval. Suppose we have a data set with a market day divided in n intervals, then we will have n seasonal models for predicting the daily volume of the day. Also, each seasonal model has two types of features: interday and intraday. The interday features corresponds to the past daily volumes and the intraday features corresponds to the past intervals volumes during the day.

3 Algorithms for Daily Volume Forecasting

3.1 Naive Strategy

As a baseline, we propose a naive strategy that uses the intraday volume information. This strategy simply sums the n prior volume intervals of the current day interval, where n is the number of total intervals of a day. As an example, in the beginning of a day, the strategy sums all intraday volume intervals of the previous day. In the end of the day, the strategy uses all the intraday information of that day. The idea behind this strategy is that the previous volume of the day is a good estimative to the volume of the next day. Moreover, this strategy reduces the volume estimative error during the day.

3.2 Partial Least Squares (PLS)

Introduced by [Wold et al., 1983] Partial Least Squares is an useful alternative to the well-know least squares method for linear regression. PLS was initially applied to analytical chemistry and is a very popular method in industries for spectroscopy in chemometrics and for quantitative structure activity relationships in drug design. Recently, PLS has been used in marketing, image processing and finances. Some of the advantages of using PLS are: good performance where the number of features are greater than the number of observations, applicable to features that are highly correlated, can be used for dimensionality reduction and prediction, easy understanding of the implicit model and fast processing.

In our work we use our PLS based in the work of [Renteria, 2003]. Suppose we have a mean centered data set $[XY]$ where X are the independent variables and Y the dependent variables. PLS extracts a latent structure using orthogonal factors as linear combinations of features. Instead of working in the features space

$$XX^T = Cov(X,X) = Var(X) \tag{1}$$

like Principal Component Analysis (PCA), PLS adds dependent variables to work in

$$XY^TY^TX = Cov(X,Y) \tag{2}$$

space. The key idea is to find a feature space that better describes both dependent and independent variables in a way that can be used for prediction of the dependent variables. The overall PLS process can be described in the following way:

```
X_1 ← X; Y_1 ← Y;
for i=1 to k do
    w_i ← X_i^T Y
    w_i ← w_i/(Y^T X X^T Y)^{1/2} //w_i normalization
    t_i ← X_i w_i
    //Computing the coefficients
    b_i ← Y_i^T t_i / t_i^T t_i
    p_i ← X_i^T t_i / t_i^T t_i
    //Residual Processing
    X_{i+1} ← X_i − t_i p_i^T
    Y_{i+1} ← Y_i − b_i t_i
end for
```

Once we have learned the PLS model we can fit to a new data set X' for predicting the Y' responses. We can do this by using w, b, p learned from the model as in the following way:

```
X'_1 ← X'; Y' ← 0;
for i=1 to h do
    //Residual Prediction
    t'_i ← X'_i w_i
    Y' ← Y' + t'_i b_i^T
    //Residual Processing
    X'_{i+1} ← X'_i + t'_i p_i^T
end for
```

3.3 Support Vector Regression (SVR)

Support Vector Machine was first introduced by Vapnik and colleagues at AT&T research lab at 90 decade within the statistical learning theory and the structural risk minimization. Their theory was proven to be very successfully on many applications of classification and function estimation. The SVM learning consists in the use of a kernel representation of the data and then formulating the problem as a convex optimization problem. Usually the convex optimization problem can be solved using a quadratic programming technique, for which the dual problem is solved. The main advantages of using SVM are: convexity of the objective function, high generalization and good performance for high dimension space of features. We must emphasize that solving a convex optimization problem implies the optimization of a quadratic function that has one local minimum.

Suppose that we have a traning data set consisting of $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \chi \times \Re$, where X are the input patterns in a higher dimension of $(\chi = \Re^d)$. The aim of SVR [Boser et al., 1992] is to find a $f(x)$ that has at most $\varepsilon-$deviation to the target y_i. The error solutions that are less than e deviations are not penalized. For example, if we are predicting a financial return function, we can establish a threshold of $\varepsilon-$loss of money. Now, suppose we have a linear function $f(x) = \langle w, x\rangle + b$, with $w \in \chi$ and $b \in R$ where $\langle ., .\rangle$ denotes the dot product in χ. For approximating $f(x)$ with an $\varepsilon-$precision, we must formulate our problem as:

$$\text{Minimize } \frac{1}{2}\| w \|^2 \tag{3}$$

$$\text{Subject to } \begin{cases} yi - \langle w, xi\rangle - b \le \varepsilon, \\ \langle w, xi\rangle + b - yi \le \varepsilon. \end{cases}$$

this formulation is good but, however, it does not allow some errors that can be irrelevant to the problem. To solve this problem , [Bennett and Mangasarian, 1992] propose the term Soft Margin that uses slack variables ξ_i, ξ_i^*. Them the formulation can be denoted as:

$$\text{Minimize } \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{Subject to } \begin{cases} yi - \langle w, x\rangle - b \le \varepsilon + \xi_i, \\ \langle w, xi\rangle + b - yi \le \varepsilon + \xi_i^*, \\ \xi, \xi_i^* \ge 0 \end{cases}$$

where $C > 0$ is the regularization constant. This problem can be solved more easily considering a dual problem utilizing Lagrange Multipliers as described in [Fletcher, 1987].

$$\text{Maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j) \\ -\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \end{cases}$$

Subject to $\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0$ and $\alpha_i,\alpha_i^*\in[0,C]$

In this dual formulation, we can see the use of kernel mapping proposed by [Boser et al., 1992] to solve nonlinear problems, where is $k(x,x')=\langle\phi(x),\phi(x')\rangle$ rather than computing $\phi(x)$ explicitly.

4 Empirical Analysis

4.1 Data Set Preparation

Our data set contains ten stocks of Bovespa which three of them consists in high liquidity stocks (PETR3, PETR4, VALE5) and the remaining are low liquidity stocks (USIM5, GOAU5, ITSA4, ITAU4, BBDC4, CSNA3). We divide a day in 28 intervals of 15 minutes and use a total of of 460 days corresponding to the dates March 2006 up to January 2008. For parameter selection, we work in the ranges March 2006 up to August 2007 and with the testing days in the range August 2007 to January 2008. For working with PLS algorithm, we preprocess the data set according to the $z-$score standardization

$$z_i=(x_i-\mu_i)/\sigma_i \tag{5}$$

where x_i represents a i-vector of a data set X,Y, μ_i is the mean of x_i and σ_i is the standard deviation. For SVR, we use the min-max normalization

$$n_i=\frac{(max_i-min_i)(x_i-\min_i)}{(cmax-cmin)+cmin} \tag{6}$$

where the values will be in the range $[cmin=-1, cmax=1]$, max_i is the maximum value of x_i, and min_i is the minimun value of x_i.

4.2 The Sliding Window Validation Method (SLW)

The key concept behind the sliding window validation is that it can convert a sequential supervised learning problem to a classical supervised learning problem. Suppose we have a data set $(x_i,y_i)_{i=1}^{N}$ of N samples where each sample is a duple (x_i,y_i). We can denote x_i as a dependent variable of y_i. In our time series prediction problem, y_i corresponds to an element to be predicted in time t and x_i can be a feature in time $t-1$ or an independent variable y_{t-1}. The SLW method maps each output y_i into an input window of features w. This input window contains a set of the w-last dependent variables disposed as $(x_{i,t-w},x_{i,t-w-1},...,x_{i,t})$. With this new sequential data set, we can apply our classical supervised learning algorithm.

Once we have the data set prepared, we must guarrantee that all the data set samples are in increasing time order. This is required to maintain a coherence with the time series problem and to exclude future observations from predicting past observations. After this process is done, we use a limited horizon window of training samples that shifts by $k-$samples per training iteration.

4.3 Error Metrics

For evaluation of our predictors we use the average of the percentage errors

$$\frac{100}{N}\sum\left(\frac{|Y_i-Y_i^p|}{Y}\right)_{i=1}^{N} \tag{7}$$

where Y_i is the true response, Y_i^p is the estimated value and N is the number of samples. Supposing we have $k-$intervals in the market day and $k-$specialized models, the average percentage error is calculated separately for each of these $k-$models.

4.4 Parameter Selection

For the parameter selection, we use a grid search heuristic. This heuristic tries values of each parameter in a search range using geometric steps. In tables 2 and 3, respectively, we can see the best parameters for PLS and SVR algorithms using the grid search with a sliding window validation method. The parameter selection data set contain examples between the period of March 2006 up to August 2007.

Stock	Avg. error (%)			Avg. num. of factors		
	1-9	10-19	20-27	1-9	10-19	20-27
PETR3	23,44	15,20	6,73	3	4	6
PETR4	17,77	10,66	3,77	3	3	4
VALE5	16,02	9,51	3,73	3	3	4
CSNA3	25,53	13,73	5,66	4	3	5
USIM5	20,16	11,70	4,83	3	3	4
BBDC4	23,22	12,31	5,54	3	3	4
ITSA4	25,07	14,99	5,95	2	3	3
ITAU4	24,84	12,28	4,86	4	3	4
GOAU4	28,33	15,99	7,40	4	3	3

Table 2. Parameter selection results for the PLS

For the SVR algorithm, we made experiments with a RBF kernel, shown in tables 7,8,9 and 10, and a default $nu=0.5$ parameter. We perform a grid search to find the best C and $Gamma$ parameter for every model of the day. The range used for C is $[2^{-5},2^{15}]$ and for $Gamma$ is $[2^{-15},2^{3}]$. We can see the best parameter selection results in the table 3. For the PLS algorithm, in table 2, we made experiments using a range $[1,30]$ of number of factors. The interesting finding is that PLS algorithm finds better results with a few number of factors. The few factors explain the

Stock	Avg. error (%)		
	1-9	10-19	20-27
PETR3	$22,49$	$14,75$	$6,83$
PETR4	$18,00$	$10,90$	3.88
VALE5	$16,48$	$9,93$	3.81
CSNA3	$24,99$	$13,81$	$5,56$
USIM5	$20,90$	$11,46$	$4,88$
BBDC4	$22,13$	$12,55$	$5,72$
ITSA4	$24,59$	$14,56$	$6,25$
ITAU4	$23,68$	$12,28$	$4,88$
GOAU4	$28,07$	$15,98$	$7,20$

Table 3. Parameter selection results for the SVR

hidden components of a financial time series such as sazonality, ciclic behaviour and tendency.

5 Forecasting Results

Table 4 presents the forecasting results when considering only the first nine intervals of the trading day. When computing the results, the average of the pencentage error for each interval is computed. Then, these errors are averaged, producing a number that indicates the system error over the considered intervals. We can see that our model benefits from the prior intraday information of the day, outperforming the naive strategy. The error reductions are significant with a 55% reduction for the best case (CSNA3) and a 19% error reduction for the worst case (USIM5). In table 4, we can also see that both PLS and SVR have competitive results in the beginning of the day (first nine intervals) with a slightly advantage for the PLS predictor. The minimum average percentage error obtained is $17,44\%$ for PETR4 stock and the maximum is $26,86\%$ for PETR3 stock. These results are very interesting because PETR3 and PETR4 stocks represents the same company.

	SVR		PLS		Naive		Best Reduction
Stock	Avg(%)	Std	Avg(%)	Std	Avg(%)	Std	$\frac{\lvert Naive-minError\rvert}{Naive}$
PETR3	$27,28$	$4,64$	$26,86$	$4,55$	$38,42$	$3,93$	$0,28$
PETR4	$18,60$	$3,10$	$17,44$	$2,59$	$22,38$	$2,73$	$0,22$
VALE5	$17,88$	$3,22$	$17,70$	$3,01$	$23,44$	$2,57$	$0,24$
CSNA3	$20,41$	$3,50$	$18,77$	$4,00$	$29,80$	$3,56$	$0,55$
USIM5	$20,28$	$2,21$	$19,75$	$1,74$	$24,64$	$2,97$	$0,19$
BBDC4	$21,85$	$3,23$	$20,34$	$3,55$	$27,97$	$2,36$	$0,27$
ITSA4	$23,32$	$3,27$	$26,23$	$3,77$	$35,61$	$3,76$	$0,34$
ITAU4	$21,65$	$3,08$	$20,99$	$3,24$	$28,66$	$3,26$	$0,26$
GOAU4	$26,23$	$4,33$	$26,65$	$3,92$	$34,47$	$3,80$	$0,23$

Table 4. Results for the initial intervals of the day (1-9)

In the middle of the market day (intervals 10 to 19), as shown in Table 5, the percentage error for all stocks is below to 18%. The best result is for the VALE5 stock, which has an average percentage error of $9,34\%$. The worst result is for the GOAU4 stock, which has a percentage error of $17,07\%$. Also, our forecasting error compared to the naive strategyr is about $33,11\%$ in average better. Note that we can see an significant improvement of the results for the middle of the day compared to the beginning of the day. There are two reasons for this improvement. One reason is that in the middle of the day the problem becomes more easy to solve. Suppose we are exactly in the middle of the day and we have to estimate the overall volume of the day. One trivial way is to sum all the previous intervals of that day and then multiply by 2. Knowing that the intraday volume of stocks has $U-$Shape pattern, we can approximate the volume of the day with the initial volume of the day. However, this strategy is not well suited, because the $U-$shape can be delayed or may not occur for some abnormal days. The second reason is that the opening of the foreign market impacts so much on the beginning of the Bovespa market day, in a way that causes much oscillation in the beginning of the day.

	SVR		PLS		Naive		Best Reduction
Stock	Avg(%)	Std	Avg(%)	Std	Avg(%)	Std	$\frac{\lvert Naive-minError\rvert}{Naive}$
PETR3	$15,77$	$1,74$	$15,43$	$1,52$	$24,41$	$3,05$	$0,36$
PETR4	$9,80$	$1,75$	$9,88$	$1,19$	$14,17$	$1,87$	$0,30$
VALE5	$9,34$	$1,35$	$9,54$	$1,27$	$14,98$	$2,08$	$0,37$
CSNA3	$11,67$	$1,56$	$10,74$	$1,23$	$19,22$	$2,39$	$0,44$
USIM5	$14,30$	$2,56$	$14,37$	$2,58$	$17,92$	$2,57$	$0,20$
BBDC4	$11,12$	$2,18$	$11,11$	$1,68$	$18,89$	$2,92$	$0,41$
ITSA4	$15,21$	$4,70$	$15,35$	$2,81$	$23,12$	$3,63$	$0,34$
ITAU4	$13,42$	$2,38$	$13,04$	$2,02$	$18,73$	$3,07$	$0,30$
GOAU4	$17,07$	$2,47$	$17,12$	$2,44$	$23,24$	$2,57$	$0,26$

Table 5. Results for the middle intervals of the day (10-19)

Finally, at the end of the market day, as shown in Table 6, our proposed model outperforms the traditional model by $45,66\%$ on average. The best result is for PETR4 with $3,93\%$ on average of percentage error and the worst result is for GOAU4 with an average percentage error of $6,97\%$. Besides the competitive results between PLS and SVR, up to the middle of the market day, curiously, SVR presents better results for all stocks in the final of the market day.

	SVR		PLS		Naive		Best Reduction
Stock	Avg(%)	Std	Avg(%)	Std	Avg(%)	Std	$\frac{\lvert Naive-minError\rvert}{Naive}$
PETR3	$6,77$	$3,20$	$6,89$	$3,19$	$14,02$	$3,92$	$0,51$
PETR4	$3,93$	$2,18$	$4,14$	$2,21$	$7,38$	$2,40$	$0,46$
VALE5	$4,16$	$2,03$	$4,19$	$2,01$	$7,58$	$2,57$	$0,45$
CSNA3	$5,03$	$2,19$	$5,17$	$2,06$	$9,17$	$3,21$	$0,45$
USIM5	$4,95$	$2,31$	$5,29$	$2,64$	$8,64$	$2,97$	$0,42$
BBDC4	$5,10$	$2,14$	$5,24$	$2,18$	$9,19$	$3,08$	$0,44$
ITSA4	$5,02$	$2,33$	$5,78$	$2,79$	$10,55$	$3,88$	$0,52$
ITAU4	$4,81$	$2,28$	$4,88$	$2,24$	$8,76$	$2,88$	$0,44$
GOAU4	$6,97$	$3,48$	$7,30$	$3,43$	$12,09$	$4,05$	$0,42$

Table 6. Results for the final intervals of the day (20-27)

In figure 2, we show an example of the forecasting error reduction during the day, for the VALE5 stock. Note that, SVR and PLS have competitive results and their behaviour are similar during the day.

6 Conclusions and Future Works

An accurate volume forecasting can be useful for tasks such as portfolio management, asset allocation and specially in algorithm trading. In algorithm trading, volume is an important characteristic of the market, especially for investors who want to minimize the market impact on their execution

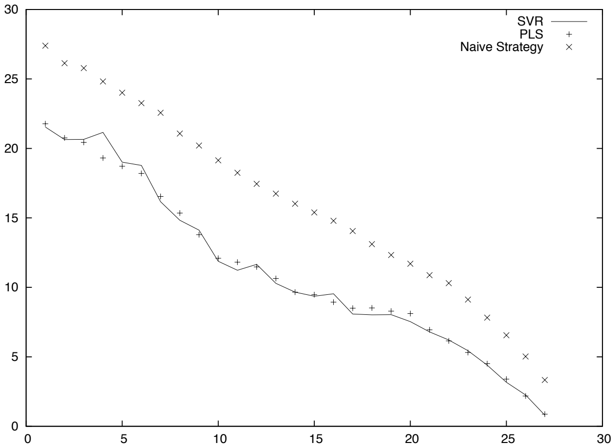

Figure 2. An example of prediction error during the day for VALE5 daily volume.

orders. Besides the great importantce of volume prediction for the market, there are few studies in financial and machine learning literature related to volume forecasting. The forecasting models used are usually for predicting partial or overall volume of a day.

In this work, we predict the overall volume of the day dealing with both overall and partial volume of the day. The idea is to use the partial volume information during the day for a better daily volume prediction. The main contribution of our work is a dynamic model for daily volume forecasting. This model updates the intraday information at each $15min-$interval. Also, we use a PLS and SVR machine learning algorithms to predict volume, differently from the work of [Bialkowski et al., 2006] that uses a Bayesian approach. Our model outperforms the naive predictor for all stocks. Mreover, our model can forecast the daily volume in the beginning, middle and end of the day, respectively, with an average percentual error of $21,26\%$, $12,89\%$ and $5,19\%$.

Our work can be used in such applications related to automated trading, portfolio management and specially in algorithm trading. In the algorithm trading, an institutional investor can use the daily volume prediction for better estimates of partial or integral volume of the day. The institutional investor can choose the daily volume estimates, in tables 4,5 and 6 that better satisfies his/her necessities to a certain threshold error. It can be done by choosing a particular percentage error that better the institutional investor's restrictions. Applications for algorithm trading usually use the volume as an additional information for a VWAP strategy.

Due to the aim of improving our single stock daily volume models, we are to extend our work to a multi-stock model. For this model, we can use correlated stocks as features or samples so as to improve a single stock prediction. An empirical analysis of stock correlations should be used in order to choose the most correlated stocks. As a possible future proposition, the predictors presented in this work (PLS and SVR) could be used as a committee so as to achieve better results. The main motivation for applying this committee of predictors would be the fact that the results of both are very competitive.

Stock	C1	g1	C2	g2	C3	g3	C4	g4	C5	g5	C6	g6
PETR3	2^{3}	$3,70.10^{-4}$	2^{1}	$4,48.10^{-3}$	2^{5}	$1,28.10^{-3}$	2^{9}	$3,10.10^{-5}$	2^{9}	$3,10.10^{-5}$	2^{7}	$1,06.10^{-4}$
PETR4	2^{5}	$1,28.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{-1}	$5,4.10^{-2}$	2^{1}	$1,56.10^{-2}$	2^{7}	$1,06.10^{-4}$	2^{1}	$5,4.10^{-2}$
VALE5	2^{7}	$1,06.10^{-4}$	2^{1}	$4,48.10^{-3}$	2^{3}	$1,56.10^{-2}$	2^{5}	$1,56.10^{-2}$	2^{5}	$4,48.10^{-3}$	2^{3}	$1,56.10^{-2}$
CSNA3	2^{13}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{7}	$4,48.10^{-3}$	2^{1}	$4,48.10^{-3}$	2^{7}	$1,06.10^{-4}$	2^{3}	$1,28.10^{-3}$
USIM5	2^{11}	$1,28.10^{-3}$	2^{-1}	$5,4.10^{-2}$	2^{1}	$1,56.10^{-2}$	2^{1}	$5,4.10^{-2}$	2^{9}	$3,10.10^{-5}$	2^{7}	$3,10.10^{-5}$
BBDC4	2^{7}	$1,06.10^{-4}$	2^{5}	$3,70.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{11}	$3,70.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{7}	$3,70.10^{-4}$
ITSA4	2^{3}	$1,28.10^{-3}$	2^{1}	$1,56.10^{-2}$	2^{1}	$1,56.10^{-2}$	2^{7}	$1,06.10^{-4}$	2^{5}	$3,70.10^{-4}$	2^{3}	$1,28.10^{-3}$
ITAU4	2^{-3}	$6,59.10^{-1}$	2^{7}	$3,10.10^{-5}$	2^{5}	$3,70.10^{-4}$	2^{11}	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$	2^{5}	$4,48.10^{-3}$
GOAU4	2^{9}	$3,10.10^{-5}$	2^{9}	$3,10.10^{-5}$	2^{1}	$5,4.10^{-2}$	2^{1}	$5,4.10^{-2}$	2^{7}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$

Table 7. SVR validation parameters for intervals 1-6

Stock	C7	g7	C8	g8	C9	g9	C10	g10	C11	g11	C12	g12	C13	g13
PETR3	2^{3}	$1,56.10^{-2}$	2^{7}	$1,28.10^{-3}$	2^{11}	$3,10.10^{-5}$	2^{11}	$1,28.10^{-3}$	2^{1}	$1,56.10^{-2}$	2^{7}	$1,28.10^{-3}$	2^{7}	$1,28.10^{-3}$
PETR4	2^{3}	$4,48.10^{-3}$	2^{3}	$1,28.10^{-3}$	2^{9}	$3,10.10^{-5}$	2^{3}	$1,28.10^{-3}$	2^{5}	$3,70.10^{-4}$	2^{5}	$3,70.10^{-4}$	2^{5}	$3,70.10^{-4}$
VALE5	2^{5}	$3,70.10^{-4}$	2^{3}	$4,48.10^{-3}$	2^{13}	$3,70.10^{-4}$	2^{5}	$4,48.10^{-3}$	2^{3}	$1,28.10^{-3}$	2^{9}	$3,10.10^{-5}$	2^{9}	$3,10.10^{-5}$
CSNA3	2^{1}	$4,48.10^{-3}$	2^{7}	$1,28.10^{-3}$	2^{9}	$1,28.10^{-3}$	2^{13}	$1,06.10^{-4}$	2^{9}	$3,70.10^{-4}$	2^{9}	$1,28.10^{-3}$	2^{3}	$1,28.10^{-3}$
USIM5	2^{5}	$1,56.10^{-2}$	2^{1}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{11}	$3,70.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{5}	$4,48.10^{-3}$
BBDC4	2^{11}	$3,70.10^{-4}$	2^{13}	$3,70.10^{-4}$	2^{9}	$3,70.10^{-4}$	2^{7}	$3,70.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{11}	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$
ITSA4	2^{7}	$1,06.10^{-4}$	2^{9}	$3,10.10^{-5}$	2^{1}	$1,56.10^{-2}$	2^{1}	$1,56.10^{-2}$	2^{5}	$1,28.10^{-3}$	2^{1}	$1,56.10^{-2}$	2^{5}	$4,48.10^{-3}$
ITAU4	2^{5}	$1,28.10^{-3}$	2^{9}	$1,28.10^{-3}$	2^{11}	$3,10.10^{-5}$	2^{9}	$1,28.10^{-3}$	2^{9}	$3,70.10^{-4}$	2^{9}	$1,28.10^{-3}$	2^{11}	$1,28.10^{-3}$
GOAU4	2^{3}	$1,56.10^{-2}$	2^{1}	$1,56.10^{-2}$	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$

Table 8. SVR validation parameters for intervals 7-13

Stock	C14	g14	C15	g15	C16	g16	C17	g17	C18	g18	C19	g19	C20	g20
PETR3	2^{9}	$1,28.10^{-3}$	2^{11}	$3,70.10^{-4}$	2^{13}	$3,70.10^{-4}$	2^{7}	$1,28.10^{-3}$	2^{1}	$1,56.10^{-2}$	2^{11}	$1,28.10^{-3}$	2^{5}	$4,48.10^{-3}$
PETR4	2^{11}	$1,28.10^{-3}$	2^{9}	$1,06.10^{-4}$	2^{9}	$1,06.10^{-4}$	2^{7}	$3,70.10^{-4}$	2^{13}	$3,10.10^{-5}$	2^{11}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$
VALE5	2^{3}	$1,28.10^{-3}$	2^{7}	$1,06.10^{-4}$	2^{9}	$1,28.10^{-3}$	2^{5}	$1,28.10^{-3}$	2^{7}	$1,06.10^{-4}$	2^{13}	$3,70.10^{-4}$	2^{7}	$3,70.10^{-4}$
CSNA3	2^{7}	$1,28.10^{-3}$	2^{9}	$3,70.10^{-4}$	2^{5}	$1,28.10^{-3}$	2^{11}	$1,06.10^{-4}$	2^{7}	$1,06.10^{-4}$	2^{3}	$4,48.10^{-3}$	2^{5}	$1,28.10^{-3}$
USIM5	2^{7}	$1,28.10^{-3}$	2^{9}	$1,06.10^{-4}$	2^{7}	$3,70.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{9}	$1,06.10^{-4}$	2^{9}	$1,06.10^{-4}$	2^{13}	$3,70.10^{-4}$
BBDC4	2^{11}	$1,06.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{13}	$3,10.10^{-5}$	2^{9}	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$
ITSA4	2^{7}	$4,48.10^{-3}$	2^{5}	$4,48.10^{-3}$	2^{5}	$1,28.10^{-3}$	2^{11}	$1,06.10^{-4}$	2^{5}	$1,28.10^{-3}$	2^{9}	$3,10.10^{-5}$	2^{9}	$3,10.10^{-5}$
ITAU4	2^{7}	$1,28.10^{-3}$	2^{9}	$3,10.10^{-5}$	2^{5}	$1,28.10^{-3}$	2^{7}	$1,06.10^{-4}$	2^{9}	$3,10.10^{-5}$	2^{11}	$3,10.10^{-5}$	2^{9}	$1,06.10^{-4}$
GOAU4	2^{3}	$4,48.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{5}	$1,28.10^{-3}$	2^{3}	$4,48.10^{-3}$	2^{5}	$1,28.10^{-3}$	2^{5}	$3,70.10^{-4}$	2^{9}	$3,10.10^{-5}$

Table 9. SVR validation parameters for intervals 14-20

7 Acknowledgements

We would like to aknowledge, firstly, CAPES and CNPQ for their support, in the form of a scolarship, which made this work possible. We would also thank my thesis advisor, which has worked togheter, both in the elaboration and revision of this work. Finally, we also thank all the people at LEARN laboratory for the suggestions.

Stock	C21	g21	C22	g22	C23	g23	C24	g24	C25	g25	C26	g26	C27	g27
PETR3	2^9	$1,28.10^{-3}$	2^{11}	$3,70.10^{-4}$	2^{13}	$3,70.10^{-4}$	2^7	$1,28.10^{-3}$	2^1	$1,56.10^{-2}$	2^{11}	$1,28.10^{-3}$	2^5	$4,48.10^{-3}$
PETR4	2^{11}	$1,06.10^{-4}$	2^{11}	$1,06.10^{-4}$	2^9	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^7	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$	2^9	$3,10.10^{-5}$
VALE5	2^{13}	$3,10.10^{-5}$	2^{13}	$1,06.10^{-4}$	2^7	$3,70.10^{-4}$	2^9	$3,10.10^{-5}$	2^{11}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^9	$1,06.10^{-4}$
CSNA3	2^5	$4,48.10^{-3}$	2^7	$3,70.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{11}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^9	$3,10.10^{-5}$	2^{13}	$3,10.10^{-5}$
USIM5	2^3	$1,28.10^{-3}$	2^1	$4,48.10^{-3}$	2^3	$4,48.10^{-3}$	2^3	$4,48.10^{-3}$	2^3	$4,48.10^{-3}$	2^{11}	$3,10.10^{-5}$	2^7	$3,70.10^{-4}$
BBDC4	2^{11}	$1,06.10^{-4}$	2^9	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$	2^{13}	$1,06.10^{-4}$	2^{11}	$1,06.10^{-4}$	2^7	$1,28.10^{-3}$	2^9	$3,70.10^{-4}$
ITSA4	2^{13}	$3,10.10^{-5}$	2^7	$3,70.10^{-4}$	2^{11}	$3,70.10^{-4}$	2^{13}	$1,06.10^{-4}$	2^{11}	$3,10.10^{-5}$	2^{13}	$3,10.10^{-5}$	2^{11}	$1,06.10^{-4}$
ITAU4	2^{13}	$3,10.10^{-5}$	2^9	$1,06.10^{-4}$	2^{13}	$3,10.10^{-5}$	2^{11}	$3,10.10^{-5}$	2^{11}	$3,10.10^{-5}$	2^9	$3,70.10^{-4}$	2^5	$1,28.10^{-3}$
GOAU4	2^7	$1,06.10^{-4}$	2^{13}	$3,70.10^{-4}$	2^5	$1,28.10^{-3}$	2^{13}	$3,10.10^{-5}$	2^5	$4,48.10^{-3}$	2^7	$1,28.10^{-3}$	2^{13}	$1,06.10^{-4}$

Table 10. SVR validation parameters for intervals 14-27

References

[Bennett and Mangasarian, 1992] Bennett, K. P. and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. Technical Report CS-TR-1991-1054a.

[Biais et al., 1995] Biais, B., Hillion, P., and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the paris bourse. *Journal of Finance*, 50(5):1655–89.

[Biakowski et al., 2005] Biakowski, J., Darolles, S., and Fol, G. L. (2005). Decomposing volume for vwap strategies. *econ.ku.dk*.

[Bialkowski et al., 2006] Bialkowski, J., Darolles, S., and Fol, G. L. (2006). Improving vwap strategies: A dynamical volume approach. *papers.ssrn.com*.

[Boser et al., 1992] Boser, B. E., Vapnik, V. N., and Guyon, I. M. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press.

[Dufour and Engle, 2000] Dufour, A. and Engle, R. F. (2000). Time and the price impact of a trade. *JOURNAL OF FINANCE*, 55:2467–2498.

[Easley and OHara, 1987] Easley, D. and OHara, M. (1987). Price; trade size, and information in securities markets. *Journal of Financial Economics*, 19.

[Engle and Russel, 1998] Engle, R. and Russel, J. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66.

[Fletcher, 1987] Fletcher, R. (1987). *Practical Methods of Optimization*, chapter 8.7 : Polynomial time algorithms, pages 183–188. John Wiley & Sons, New York, second edition.

[Gouriroux and Fol, 1998] Gouriroux, C. and Fol, G. L. (1998). Effet des modes de ngociation sur les echanges, revue economique. *Revue Economique*, 49.

[Lean et al., 2008] Lean, Y., Shouyang, W., and LAI, K. (2008). Forecasting china's foreign trade volume with a kernel-based hybrid econometric-ai ensemble learning. *:* .

[Lux and Kaizoji, 2007] Lux, T. and Kaizoji, T. (2007). Forecasting volatility and volume in the tokyo stock market: Long memory, fractality regime switching. *Journal of Economic Dynamics and Control*.

[McCulloch, 2004] McCulloch, J. (2004). Relative intraday volume on the nyse modelled as a doubly stochastic binomial point process. *Working Papers Series*.

[Renteria, 2003] Renteria, R. (2003). *Algoritmos para Regressão por Mínimos Quadrados Parciais*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

[Wold et al., 1983] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method. In *Proceedings Conf. Matrix Pencils*, pages 286–293. Springer Verlag.