

BIG DATA ANALYSIS OF FUTURES: PENALIZED REGRESSION SPLINES
FOR TRADE VOLUME PREDICTION AND PRICE VOLATILITY VS TRADE
VOLUME RELATIONSHIP

by

Aniver Oluwatobi Bosede

A thesis submitted to Johns Hopkins University in conformity with the
requirements for the degree of Master of Science in Engineering

Baltimore, Maryland

© 2017 Aniver Oluwatobi Bosede

All Rights Reserved

Abstract

Acknowledgments

Contents

Abstract	ii
Acknowledgments	iii
List of Tables	v
List of Figures	vi
Introduction	1
Methods	3
Results	9
Discussion	9
Appendices	9
Bibliography	10
Curriculum Vitae	12

List of Tables

Predictors used in models	8
Dummy table	9

List of Figures

Histograms of before and after Box-Cox transformation of trade volume . .	4
Plot of trade volume as days pass and as time to maturity gets further. . .	4
Histograms for instrument types	5
Plot of trade volume as days pass for instrument types	6
Plot of trade volume as time to maturity gets further for instrument types	7
Dummy figure	9
Dummy figure	9

Introduction

Technology has advanced so far that in our society today we are constantly collecting data [1]. This has created the issue of how to feasibly analyze such overwhelming amounts of data [1]. We can then consider how to efficiently store it or how to best carry out statistical computations [9]. The aforementioned together form the relatively new and evolving field of big data. The research here focuses on the analysis of hundreds of million of rows of futures trading data with the aid of Apache Spark (Spark).

Spark is a fault-tolerant and general-purpose cluster computing system providing APIs in Java, Scala, Python, and R [5]. Spark was chosen for the analysis over the similarly popular Hadoop MapReduce (MapReduce) because of Spark's performance advantages as well as its greater computational capabilities [8]. Not only does Spark cache data resulting in persisted in-memory manipulations [7], it also includes Structured Query Language (SQL) and MLlib, a machine learning library [5]. Conversely, MapReduce only does manipulations via disk reads and thus does not allow for data sharing [9]. Furthermore for a typical pipeline, external systems would have to be combined with MapReduce to provide querying and machine learning functionality [9]. In this case, A stand alone set up was employed, meaning that analysis was done using a one node cluster setup or one machine. Out of convenience the local file system was used for storage as opposed to a database like Cassandra or HDFS (Hadoop Distributed File System). However, for long term data analysis, it would be worthwhile to invest the time needed to set up a more robust data storage system.

The futures trading data come from the Chicago Mercantile Exchange (CME) and were collected from May 2, 2016 to November 18, 2016. Raw data from the CME included extended hours trading and was collected via the Trading Technologies X_TRADER® API RTD (Real Time Data) server. The server returned raw records with instrument name, maturity, date, time stamp, price, and quantity fields. The

futures were comprised of 21 financial instruments spanning six markets - foreign exchange, metal, energy, index, bond, and agriculture recording roughly a trade every half second. First, this work uses spline regression to predict the volume of trading for any given day. Volume during a particular time period is taken to mean the number of units traded. A spline regression was chosen due to lack of knowledge regarding the likely non-linear function underlying the response to covariates, in particular time to maturity. Predicting trade volume is of interest because many trading algorithms depend on volume [6]. Additionally, accurate volume predictions over a given interval allows traders to be more effective [6]. In general, volume prediction increases trading strategy capacity, controls trading risk, and manages slippage [6].

Second, the relationship between price volatility and trade volume is explored using standard deviation as a measure of volatility. In particular, volatility vs daily volume and volatility versus hourly volume were plotted to see whether or not the correlation remained the same with the passing of days and hours. It should be noted that unpredictable volume shocks have been known to be more predictive of change in volatility than predictable volume changes [2]. This volatility-volume relationship is of importance due to the notion that hedgers are motivated to trade futures to stabilize their future income flows or costs, wherein the volume of their trading is based on their expectation of price variability [3]. Likewise, speculators are motivated to trade futures based on expectations of price variability [3]. Due to the fact that new information on the market causes agents such as hedgers and speculators to trade until prices reach a revised equilibrium which then changes price and trading volume, we expect a positive correlation between volatility and volume [3]. Indeed past research indicates that there is a positive relationship between volume and price volatility [3]. This sort of exploration provides information on the efficiency of futures markets which regulators can then use to decide upon market restrictions [3].

Methods

A cubic regression spline was thought to be appropriate for modeling trade volume. Spline regression derives its name from a draftsman’s spline which is a flexible strip of metal or rubber used to draw curves [4]. Similarly, spline basis functions are piecewise polynomials used in fitting curves which are linear in terms of the basis function. Splines have been used, principally in the physical sciences as well as in biomedicine, to approximate a wide variety of functions [4]. Cubic splines in particular have been found to have nice properties with good ability to fit nonlinear curves. Cubic splines can be made to be smooth at the knots, endpoints of intervals on the x-axis, by forcing the first and second derivatives of the function to agree at the knot [4].

Holidays were removed from the raw data. Then the day of the month, day of the week, and hour of the trade were extracted from the time stamp. An aggregation was then done to sum the number of trades per hour for each product, where product is defined as an instrument-maturity pair. There were 148 such products. Aggregation reduced the data from 105 million records to 8826 records. Day, time to maturity, and market fields were created and total trade volume for each day was calculated.

Exploratory analysis was then done on the reduced data set. To ensure that a spline regression was appropriate for modeling trade volume, the first thing done was to create histograms of the trade volumes. One of the assumptions behind regression is that the response conditioned on the predictors is normally distributed. Even if normality fails, regression is feasible, but under normality least squares is optimal. Thus transforming to normality is desirable. The histogram of the raw trade volume was skewed as shown in left of Figure 1. Therefore the volume was Box Cox transformed, after which the data became normal as shown in right of Figure 1.

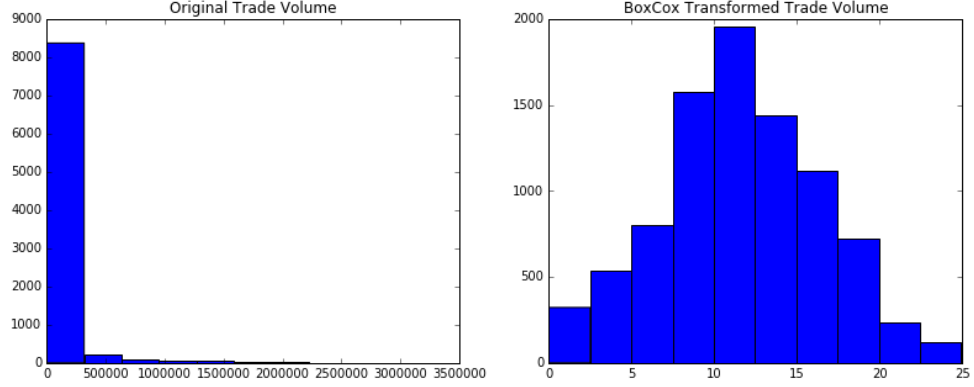


Figure 1: Histograms of before and after Box-Cox transformation of trade volume

There was curiosity regarding how trade volume changed as time passed. Thus a plot of the transformed trade volume versus day (ignoring weekends) was made. Time to maturity was not taken into consideration for instruments to reduce the number of points plotted. The aggregated trade counts on just instrument reduces the data to 2789 rows. However this resulted in the sums of trades increasing, since all trades of the same instrument in the hour were counted regardless of maturity. Referring to the left plot in Figure 2, the trade volume appears constant across the days with a cluster of high volume trades above 55 which is 797,995 trades. and a cluster of low volume trades below 45 which is 282,640 trades. This implies medium trading is be-

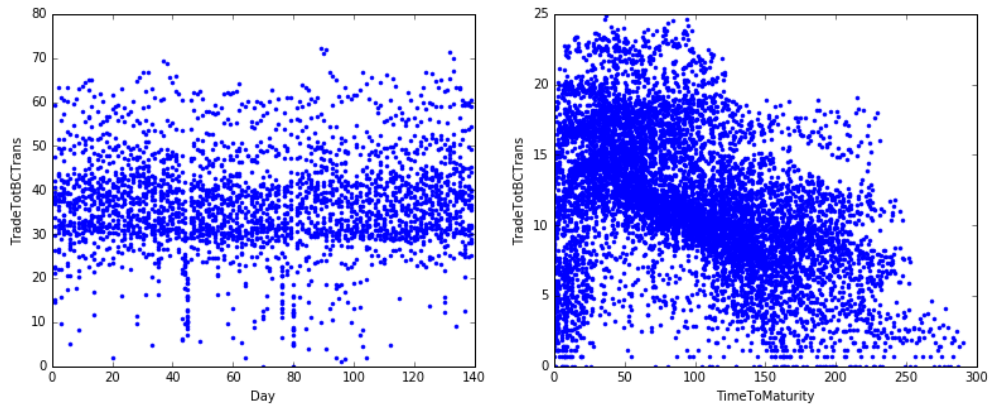


Figure 2: Plot of trade volume as days pass and as time to maturity gets further.

tween 282,640 and 797,995. Also intuitively, it makes sense that less trading occurs far from maturity and near maturity. Far from maturity speculators might not have any information that would move them to purchase a future and most hedgers may only seek to minimize risk in the short term. Then near maturity traders are closing their positions. To confirm this theory a plot was made of the transformed trade volume versus time to maturity, right in Figure 2. Looking at the plot it appears that the data follows intuition.

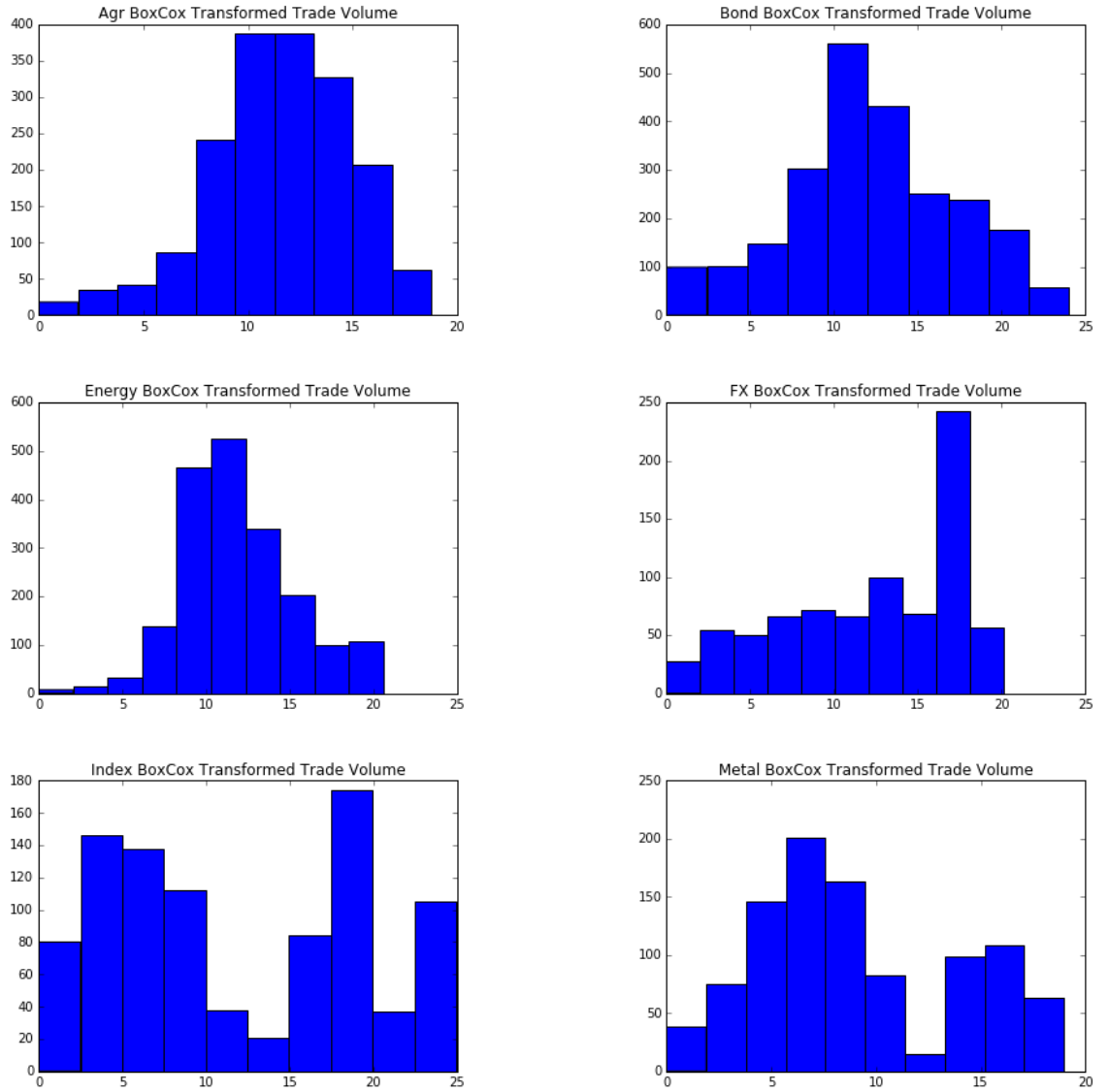


Figure 3: Histograms for instrument types

Whether it made sense to make one model or several models for each instrument market - foreign exchange, metal, energy, index, bond, and agriculture was also considered. This would mean that each of the markets need to have near normal distributions. In Figure 3 it can be seen that agriculture, bond, and energy are approximately normally distributed after Box-Cox transformation, but foreign exchange, index, and

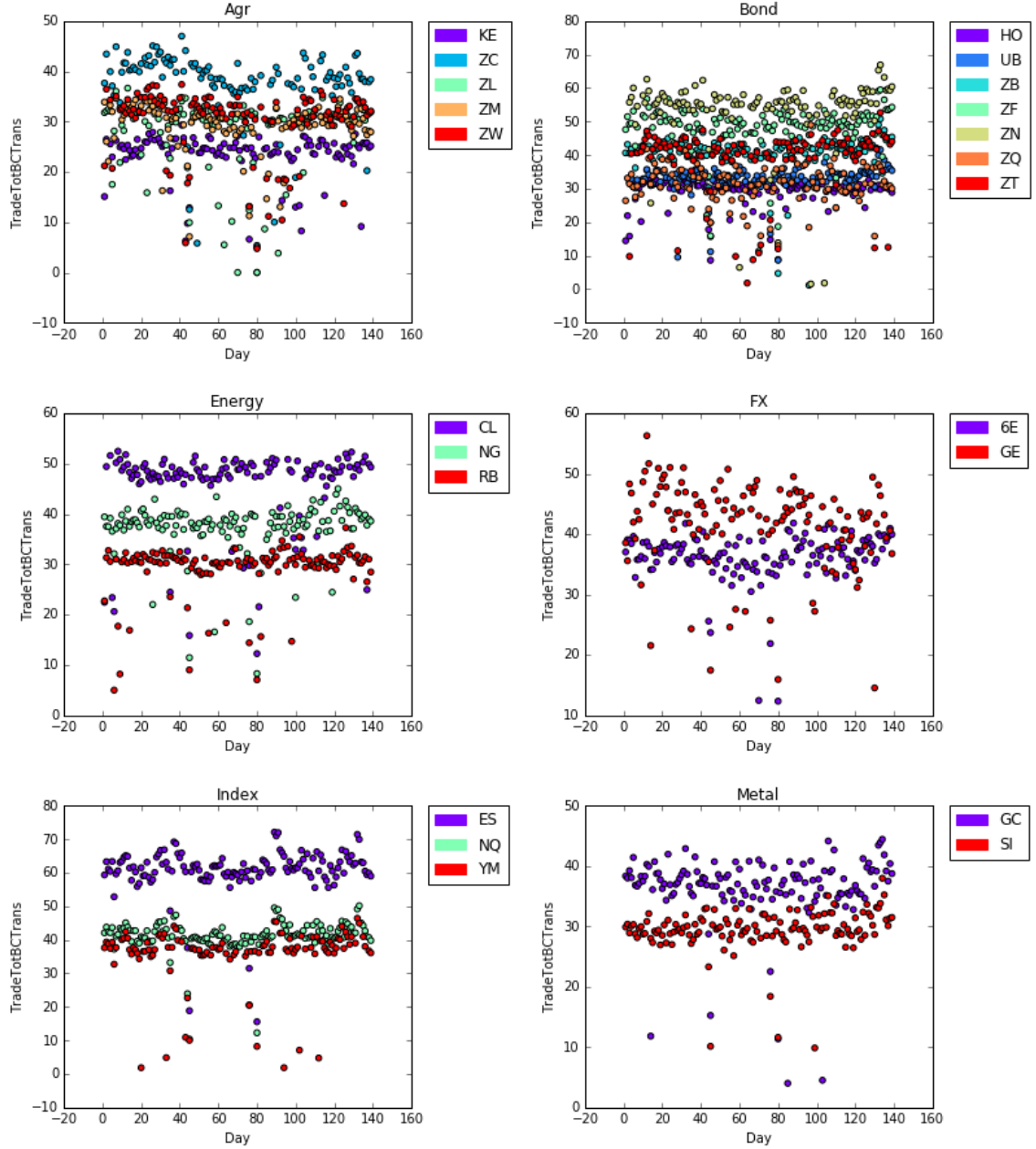


Figure 4: Plot of trade volume as days pass for instrument types

metal are not. Therefore we expect higher error in the spline model of the latter compared to the former group of instruments. The trend in daily trade volume as time passed was also explored for each instrument market in Figure 4. These points were then color-coded to understand which instruments comprised the various clusters. It is clear that low volumes of trading occur for metal and agriculture markets compared with bond and index which appear to be more liquid.

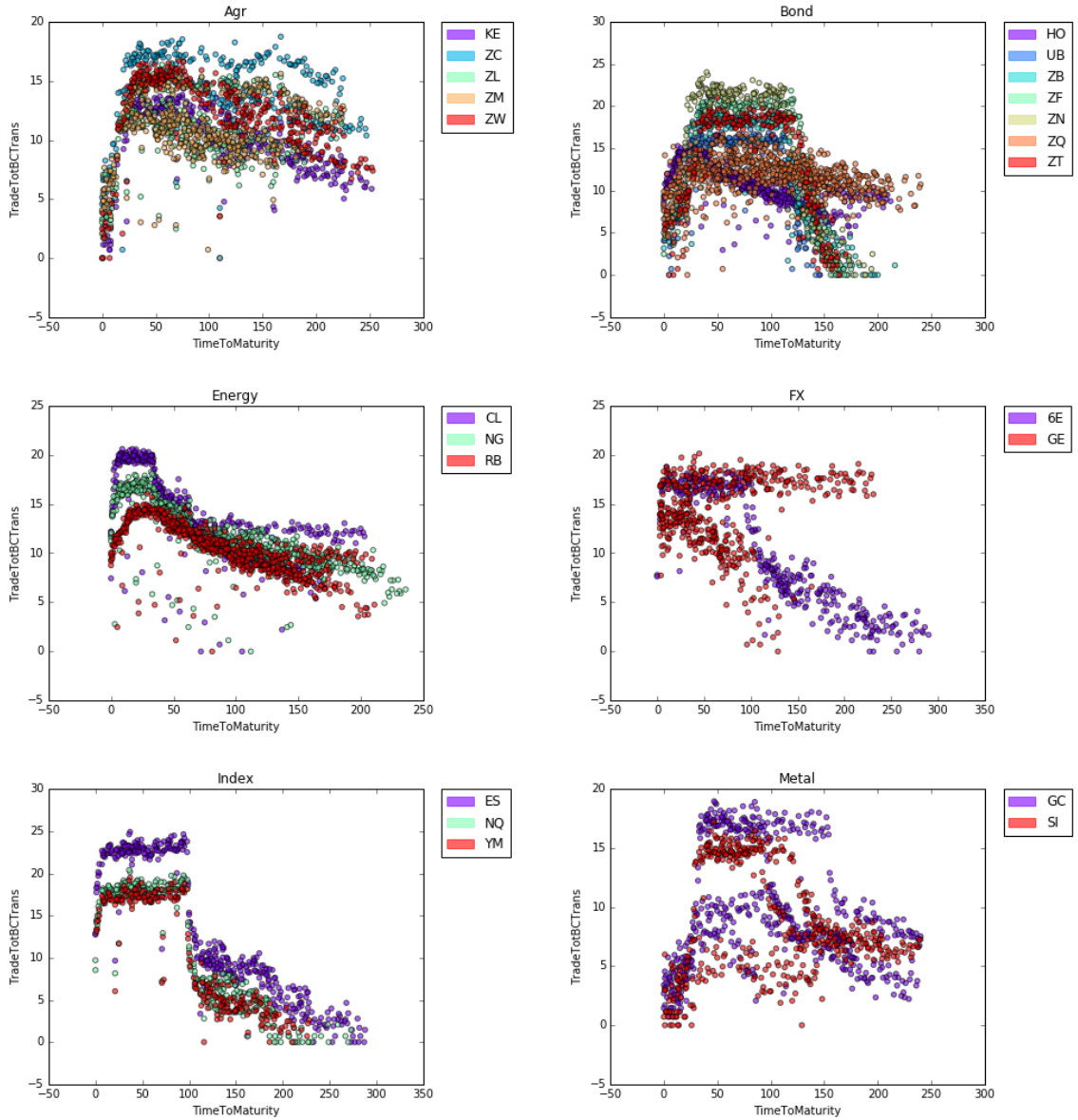


Figure 5: Plot of trade volume as time to maturity gets further for instrument types

Finally the trend in daily trade volume as time to maturity increased was visualized for each market in Figure 5. Like the histograms, agriculture, bond, and energy plots follow the general trend of low trade volume far from the maturity and close to maturity, but foreign exchange, index, and metal do not. There is some clustering in the metal and index plots as well which seems in line with the bimodal and trimodal histograms respectively.

The above described exploratory analysis deemed spline regression reasonable for trade volume prediction. Thus, both one linear model and multiple linear models paradigms were considered, two versions for each, making four models in total as shown in Table 1. All models under consideration were linear. One model paradigm meant a single model for all markets, whereas the multiple models paradigm meant each market had its own model. Spline variables are denoted by “s”. The models were fit on May to August data and used for forecasting trade volume from September through November. The mean absolute deviation (MAD) was then calculated for each model to compare the errors of forecasted volumes in a robust manner. MAD is defined as: $\sum_{i=1}^N \frac{|\hat{Y}_i - Y_i|}{N}$. The models were penalized with an integrated square second derivative cubic spline. This amounted to a natural spline and so generalized cross validation was employed to find an optimal smoothing parameter. The knots were placed at fixed intervals.

Table 1: Predictors used in models

One Linear Model		Multiple Linear Models	
model 1'	model 1	model 2'	model 2
s(TimeToMaturity) s(DayofMonth) Market DayOfTheWeek	s(TimeToMaturity) s(DayofMonth) Market DayOfTheWeek Instrument	s(TimeToMaturity) s(DayofMonth) DayOfTheWeek	s(TimeToMaturity) s(DayofMonth) DayOfTheWeek Instrument

An additional goal of this study was to understand the relationship between price

volatility and trade volume. To this end, hourly aggregation similar to that which was done for trade volume was done for price volatility. Volatility was measured using standard deviation. Next each product's hourly price volatility was plotted against its hourly trade volume for the full period over which the data was collected; that is from May to November. Given past research, we expect a positive correlation between price and volume. Consistency of the correlation between price and volume over the days was of interest as well. This was investigated for each product by plotting daily correlations between hourly price volatility and hourly trade volume.

Results

Figure 6: Dummy figure

Figure 7: Dummy figure

Table 2: Dummy table

Discussion

Appendices

Bibliography

- [1] The data deluge, Feb 2010.
- [2] Hendrik Bessembinder and Paul J Seguin. Price volatility, trading volume, and market depth: Evidence from futures markets. *Journal of financial and Quantitative Analysis*, 28(01):21–39, 1993.
- [3] Andrew J Foster. Volume-volatility relationships for crude oil futures markets. *Journal of Futures Markets*, 15(8):929–951, 1995.
- [4] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- [5] Xiangrui Meng, Joseph Bradley, B Yuvaz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, D Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *JMLR*, 17(34):1–7, 2016.
- [6] Venkatesh Satish, Abhay Saxena, and Max Palmer. Predicting intraday trading volume and volume percentages. *The Journal of Trading*, 9(3):15–25, 2014.
- [7] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association, 2012.
- [8] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. *HotCloud*, 10:10–10, 2010.
- [9] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman,

Michael J Franklin, et al. Apache spark: a unified engine for big data processing.
Communications of the ACM, 59(11):56–65, 2016.

Curriculum Vitae