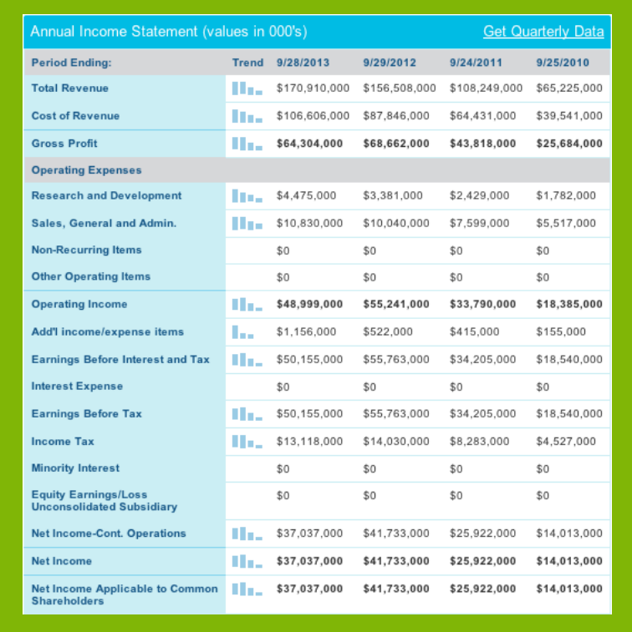
Predictive Modeling: Financial Momentum

By: Zach Freedman, Advised By: Takunari Miyazaki
Trinity College Computer Science Department

Abstract

How does a company's prior performance affect it's future share price? The aim of this project is to see whether a company's previous financial success or failure carries over into the next performance period.

To discover whether such a correlation exists, this project analyzes companies based off of their previous financial statements, in addition to their pricing history.



Annual Income Statement (values in 000's)					Get Quarterly Data
Period Ending:	Trend	9/28/2013	9/29/2012	9/24/2011	9/25/2010
Total Revenue		$170,910,000	$156,508,000	$108,249,000	$65,225,000
Cost of Revenue		$106,606,000	$87,846,000	$64,431,000	$39,541,000
Gross Profit		$64,304,000	$68,662,000	$43,818,000	$25,684,000
Operating Expenses					
Research and Development		$4,475,000	$3,381,000	$2,429,000	$1,782,000
Sales, General and Admin.		$10,830,000	$10,040,000	$7,599,000	$5,517,000
Non-Recurring Items		$0	$0	$0	$0
Other Operating Items		$0	$0	$0	$0
Operating Income		$48,999,000	$55,241,000	$33,790,000	$18,385,000
Add'l income/expense items		$1,156,000	$522,000	$415,000	$155,000
Earnings Before Interest and Tax		$50,155,000	$55,763,000	$34,205,000	$18,540,000
Interest Expense		$0	$0	$0	$0
Earnings Before Tax		$50,155,000	$55,763,000	$34,205,000	$18,540,000
Income Tax		$13,118,000	$14,030,000	$8,283,000	$4,527,000
Minority Interest		$0	$0	$0	$0
Equity Earnings/Loss Unconsolidated Subsidiary		$0	$0	$0	$0
Net Income-Cont. Operations		$37,037,000	$41,733,000	$25,922,000	$14,013,000
Net Income		$37,037,000	$41,733,000	$25,922,000	$14,013,000
Net Income Applicable to Common Shareholders		$37,037,000	$41,733,000	$25,922,000	$14,013,000

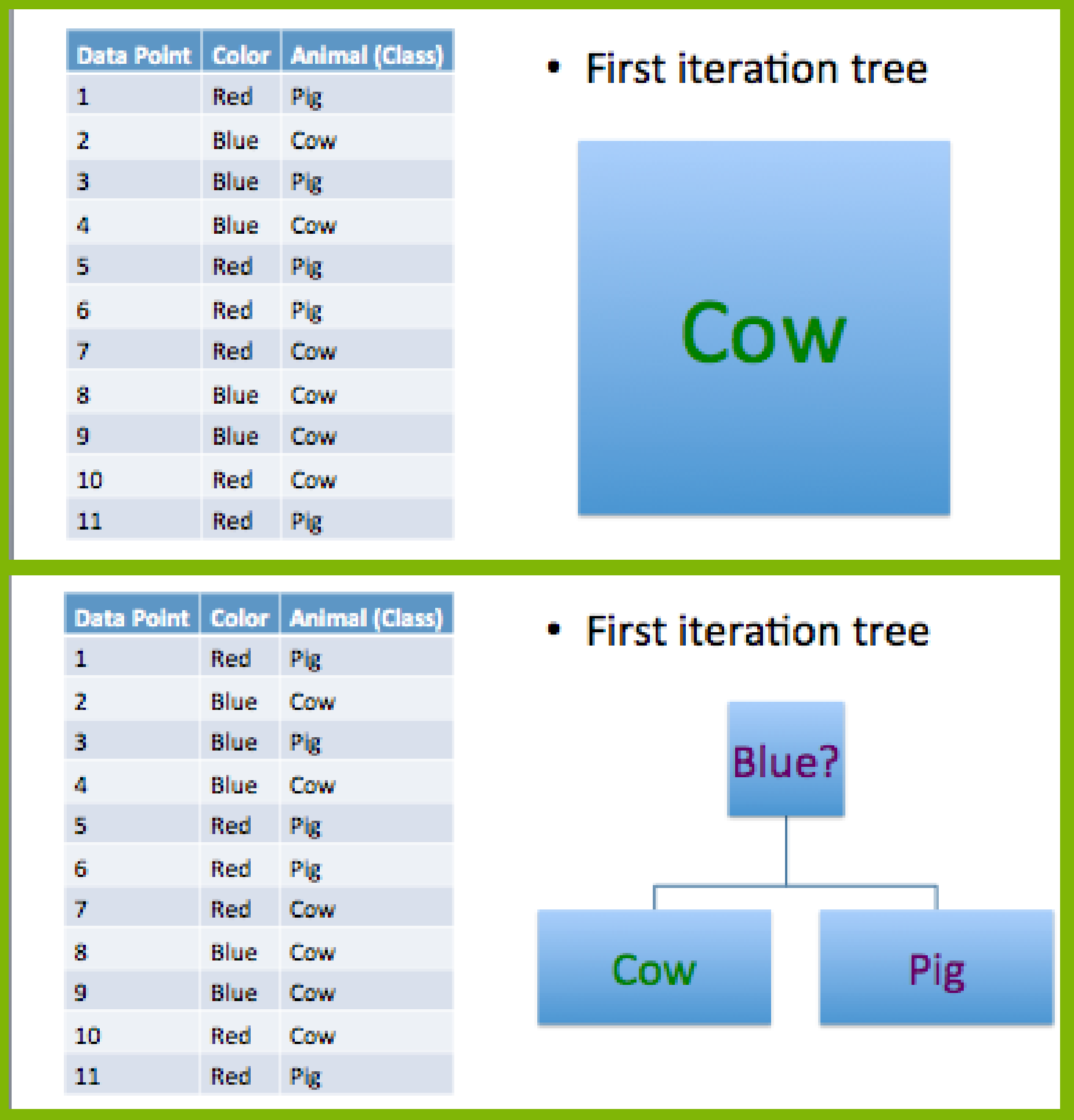
Figure 1

4 Year
04/07/14
$760.00
680.00
600.00
520.00
480.00
440.00
400.00
360.00
320.00
280.00
240.00
200.00
Jul
Oct
11
Apr
Jul
Oct
12
Apr
Jul
Oct
13
Apr
Jul
Oct
14
Apr

Figure 2

Companies are classified based on their performance. A company that loses money should be classified as a "Sell", whereas a company that gains money should be a "Buy". These classifications are made using decision trees, which classify data objects into predefined classes based on attribute tests.

Decision Tree Basics



Data Point	Color	Animal (Class)
1	Red	Pig
2	Blue	Cow
3	Blue	Pig
4	Blue	Cow
5	Red	Pig
6	Red	Pig
7	Red	Cow
8	Blue	Cow
9	Blue	Cow
10	Red	Cow
11	Red	Pig

- First iteration tree

Data Point	Color	Animal (Class)
1	Red	Pig
2	Blue	Cow
3	Blue	Pig
4	Blue	Cow
5	Red	Pig
6	Red	Pig
7	Red	Cow
8	Blue	Cow
9	Blue	Cow
10	Red	Cow
11	Red	Pig

- First iteration tree

Figure 3

Introduction

To discover whether prior performance affects future growth for financial entities, similar to the concept of the conservation of momentum in physics, the following data pipeline is followed:

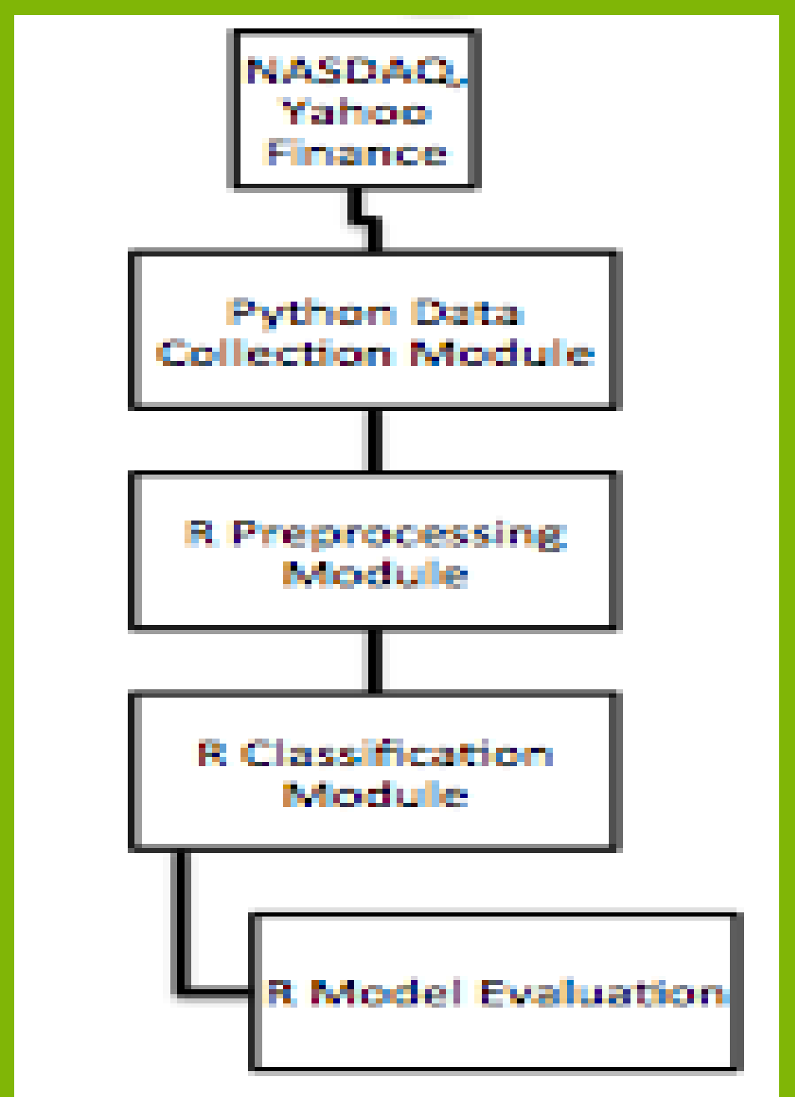


Figure 4

Data is first scraped from NASDAQ and Yahoo Finance using a scraper written in Python which implements Beautiful Soup, a popular HTML processing package. This scraping process creates the database.

After the database has been populated, the preprocessing module, written in R, removes noisy data objects and performs dimensionality reduction on the data set to simplify both the algorithmic complexity and the difficulty of the predictive model growth.

After the data has been preprocessed and written to a new .csv file, which represents the refined database, the new database is fed into the predictive classification module, where the decision tree is grown and its performance is evaluated.

Tools

Beautiful Soup An HTML parsing package implemented in Python that allows users to process HTML code.

Python An open source, general-purpose programming language

R A statistical programming language, providing users with ready to use graphical interfaces and data mining packages

Modules

Data Collection This module is responsible for database population. Data was not readily available, so it needed to be obtained.

Two scrapers are written in Python using Beautiful Soup to obtain the required data. One collects financial statements from NYSE and NASDAQ. The other obtains pricing history from Yahoo Finance.

Data Preprocessing The preprocessing module eliminates noisy data objects while also reducing the dimensionality of the dataset. Data objects missing attributes, such as an annual report for a given period, or with irregular behavior, such as pricing history containing share splitting, is eliminated from the dataset.

Additionally, the attributes representing a financial entity are significantly reduced to a feature subset selection vector, where only the most essential attributes summarizing financial growth are present.

Predictive Modeling Decision tree growth occurs in this final module of the package. After reading in the new database populated above, a decision tree is grown from a sample of the dataset to avoid overfitting.

The decision tree is grown using the pre implemented R package, *rpart*, which is the Recursive Partitioning and Regression Trees package in the R library. After the tree is built, it can be plotted for visualization and also evaluated based on predictive accuracy.

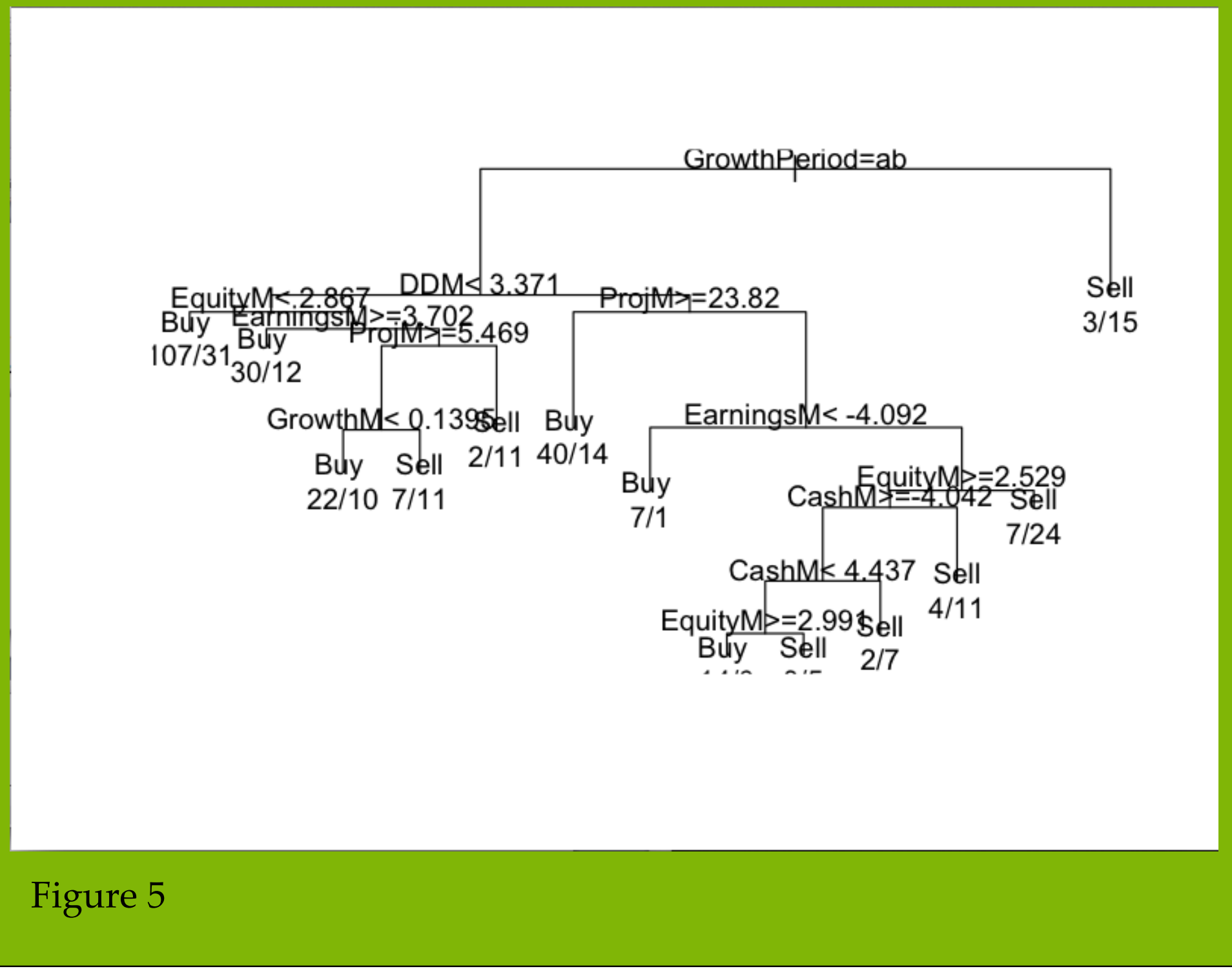


Figure 5

Results

There were two main classification sets used for the data. One set consisted of three possible classifications (Strong Buy, Buy, and Sell), while the consisted of only two (Buy and Sell). On average, the decision tree produced from the ternary classification set, using a training set roughly one tenth the size of the entire dataset, produces a classification accuracy of 42%.

The binary classification set produces an average predictive accuracy of roughly 60%, exceeding the outcome of randomly guessing buy or sell by 10%, but also falling short of the expected accuracy: 70%.

The summary of the tree's predictions is given in matrix form, such as in **Figure 5** below.

	Buy-P	Sell-P
Buy-A	0.383	0.233
Sell-A	0.187	0.197

Figure 6

Conclusions

From the results shown from the predictive model, it's reasonable to say that, by only looking at a company's financial statements and pricing history, we cannot improve our predictions much more than the evenly distributed binary probability of 50/50, or in the case of the ternary classification, 1/3 for each class. However, by performing attribute tests on the data (as done with the decision tree), we can increase the predictive accuracy to 1.2 times the original random choice accuracy.

For our ternary model, this means 33*1.2 = 39.6%, whereas for our binary model, this yields 50*1.2 = 60%.

Acknowledgements

- Trinity College Computer Science Department

- "NASDAQ Stock Market." <http://www.nasdaq.com>
- Lattanzi, R. (2012, September) Stock Selection using non-linear classifier: a Decision Tree based approach, CCFEA
- Huerta, R. (2013) Nonlinear support vector machines can systematically identify stocks with high and low future returns, Algorithmic Finance, 2, 45-58
- Cont, R. (2010, April 08) Empirical properties of asset returns: stylized facts and statistical issues, Quantitative Finance, 1:2, 223-236