Algorithms for merging tick data and data analysis for Indian financial market

Sinha, Pankaj and Sharma, Gopalakrishna and Shah, Akash and Singh, Abhijeet

Faculty of Management Studies, University of Delhi

3 July 2011

Algorithms for Merging Tick data and Data Analysis for Indian Financial Market

Pankaj Sinha
Faculty of Management Studies, University of Delhi

P. Gopalakrishna Sharma
Indian Institute of Technology, Guwahati

Akash Shah
Indian Institute of Technology Kharagpur

Abhijeet Singh
Institute of Technology, Benaras Hindu University

Abstract

This paper discusses the problem of '**merging**' financial tick data available from data sources such as Bloomberg, NSE, and Thomson Reuters etc. Different derivative securities are traded on the exchange with different frequencies in each unit of time such as second or minute in intraday trading , therefore, it is difficult to form 'ordered pairs', which are essential for any financial analysis, of tick data representing the simultaneous trades of the different derivative securities. Merging refers to the conversion of intraday tick data of different securities of varying frequencies, as provided by data sources, into the form in which the tick data of all traded derivative securities have same frequency, so that it is possible to form ordered pairs of data (in every unit time period) in such a way that the original nature of the data is preserved.

The four merging algorithms: Truncation, Weighted mean, median and all-combinations algorithm are compared with *Dropdown algorithm*, which is being used widely by the trading firms. Using NSE intraday tick data for various trading days, it is found that 'Truncation' and 'Weighted Mean' algorithms are more efficient merging algorithms.

1. Introduction

In recent times, the growth of world financial markets has been exponential both in size and exposure. With increasing globalization, market and company developments require rapid, accurate, and sophisticated in-depth analysis. In order to enhance the analysis capacity of financial data and to work on live data feed, a Financial Lab is required. The main objectives of Finance lab is to give efficient tools to interpret the financial market, to enable rigorous and efficient research with the financial data as well as to provide a basis for validating trading and investment strategies, 'without involving real money, but using real, live data'. In Indian Context, getting live data (without any kind of lag) from five financial markets in India 'NSE Cash, NSE F&O, MCX, BSE Cash and NCDEX' is an essential component of finance lab. All the features and options provided in these terminals are exact replicas of what the brokers use.

The following resources required in Finance Lab as in 'Rotman School of Management, University of Toronto' are:

Equipment

- Dual flat screen workstations (specs)
- Bloomberg terminal
- Projection screens
- Integrated sound system with remote microphone
- Data walls

Data Feeds

- Bloomberg
- Reuters 3000 Xtra
- Reuters Station

Research Databases

- Compustat
- CRSP
- Datastream
- ExecuComp
- Financial Post
- IBES
- ISSM
- LIFFE
- Mergent FISD
- OptionMetrics Ivy DB
- Reuters DealScan
- Reuters LoanConnector

- SDC Platinum Global New Issues
- SNL Financial
- TAQ
- TSX(CFMRC)
- TSX Tick-by-Tick Market Data
- Worldscope

FRTL Documents

- Alternative Equities Helpfile

Data Access

- Wharton Research Data Services (WRDS)
- FRTL Fileserver
- CHASS Datacenter

Applications

- Rotman Interactive Trader
- Rotman Portfolio Manager
- Portfolio Chooser
- Capital IQ

Data lies at the crux of the finance lab. Though financial data may be easily available in a finance lab, but the data seems to be unfit from the perspective of statistical and financial analysis. Since different derivative securities traded on the exchange have different frequencies in every unit time period (such as second or minute) in intraday trading, therefore, it is not possible to form 'ordered pairs' of tick data representing the simultaneous trades of the different derivative securities. For any financial analysis, it is required that the intraday tick-data is merged in such a way that it is possible to form ordered pairs of data columns in each unit time while preserving the original nature of the data. Before providing the various algorithms for merging financial data, we need to take a closer look at Indian Financial market and type of data available from them.

2. Markets and Data

Bombay Stock Exchange (BSE) was the first among all twenty-two exchanges existing in India. Having begun formal trading in 1875, it is one of the oldest in Asia. In the past few years, many big changes in the Indian securities market have been witnessed, especially in the secondary market. The key features of India's equity market are; firstly, electronic limit order books, which have become the norm nowadays with all Exchanges based on them. In addition to this, there has been an increase in the level of information flow and awareness among the traders, leading to an integration of markets all over the country.

Among the processes that have already started are electronic settlement trade and exchange-traded derivatives. Before 1995, an open-outcry system was used in India, a process in which traders shouted and hand-signaled from one single platform. All exchanges started shifting to screen-based trading around 1993, motivated because of a need felt for greater transparency. National Stock Exchange (NSE) was the first Exchange to shift to an electronic order-book system.

Before 1994, India's stock markets were dominated by BSE. But due to inefficient communication arising out of lack of technology, there was a price differentiation arising between markets in Mumbai and other parts of the country. Such price-differencing errors limited order flow to markets within Mumbai. "Explicit nationwide connectivity and implicit movement toward one national market has changed this situation" (Shah and Thomas, 1997). NSE has created established satellite communication systems, allowing equal access to all members of the Exchange. Soon after, BSE and the Delhi Stock Exchange also increased their respective numbers of trading terminals all over the country, with the presence of arbitrages eliminating pricing discrepancies between markets. Five markets are presently most active in India: BSE, NSE F&O, NSE Cash, MCX and NCDEX.

According to the annual survey of derivatives exchanges for the International Options Markets Association (IOMA) covering 54 exchanges in 2010 (conducted by the World Federation of Exchanges), "with a 26% growth rate in 2010, trading in derivatives contracts on regulated exchanges worldwide surged to its highest level since 2004. 22.4 billion derivative contracts were traded on exchanges worldwide (11.2 billion futures and 11.1 billion options) against 17.8 billion in 2009. The highest growth rate (+41%) was observed in the Asia - Pacific region and for the first time volumes traded in this geographical zone were higher than on the two other zones."

It goes ahead to state that the NSE turned out to be the second largest Exchange in the world, based on the number of stock index options contracts traded in 2010, and fourth largest Exchange based on the number of stock index futures contracts traded in 2010.

Index Futures were accounting for half of the nominal value of equity derivatives traded in 2010. The bigger size of those contracts shows that they are more often used by institutional investors for hedging purposes, although the work done by us attempts to highlight the nature of the data from an arbitrage-based point of view. In 2010 volumes increased slower (+5%) than

other derivatives markets. As seen in Figure-1, growth of index options is faster than that of index futures since the last few years.

"In Asia, the market is much less concentrated. Five Asian Exchanges are represented in the list of ten most active exchanges in the world for volumes of index futures traded: Hong Kong Exchanges, Korea Exchange, National Stock Exchange of India, Osaka SE and Singapore Exchange. Increase in Osaka and Singapore was offset by the decrease perceived in India."

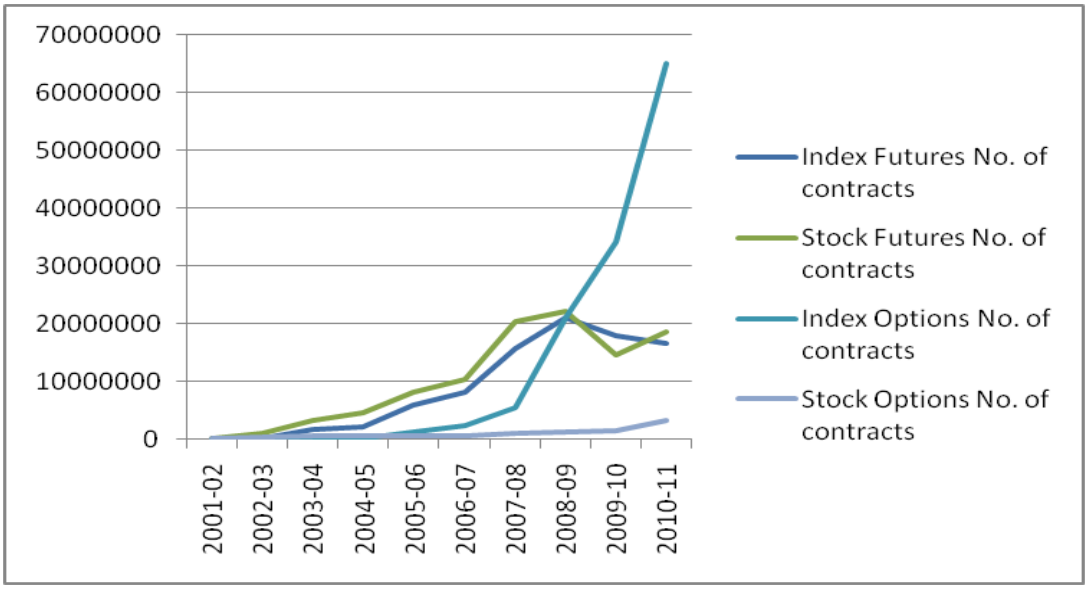

Figure 1: Volumes of different Instruments.

Having considered the different markets in India, we can safely ascertain that the NSE is the exchange with the maximum amount of volumes traded. Henceforth, all the data analysis that we are going to present will be pertaining to the NSE itself.

There are three kinds of data that are available for/from the NSE:

- End of day data
- Minute data
- Tick-data

End of Day Data:

End of Day data is available on the NSE website, www.nseindia.com; minute data was made available on BLOOMBERG, while tick-data was made available on CDs purchased from NSE.

End of Day data for any derivative is the weighted mean for that derivative calculated for the last thirty minutes of trading for each trading day. It is freely available on the NSE website

for each day. Another advantage of analyzing end of day data, besides the free availability of the same, is that it can be used to analyze the data from a long term perspective, over a period that can be as long as two years. However, one disadvantage is that, this type of data is available only at the end of the day, and as such, no data can be made available as the market varies. Hence, the variations which happen within the duration of one trading day cannot be capitalized upon.

A sample of end of the day data of NIFTY future of expiry 26/05 as provided by NSE is given in Table-1. It only reports intraday low and high price as well as daily opening and closing prices.

Symbol	Date	Expiry	Open	High	Low	Close	LTP	Settle Price	No. of contracts	Turnover in ₹ Lacs	Open Int	Change in OI	Underlying Value
NIFTY	02-May-2011	26-May-2011	5769.00	5790.00	5703.15	5725.40	5719.10	5725.40	301796	865444.99	24566250	936500	5701.30
NIFTY	03-May-2011	26-May-2011	5698.70	5726.00	5560.00	5568.70	5562.25	5568.70	534463	1507516.59	26546850	1980600	5565.25
NIFTY	04-May-2011	26-May-2011	5541.30	5592.90	5510.65	5538.05	5536.00	5538.05	442230	1225589.60	26687350	140500	5537.15
NIFTY	05-May-2011	26-May-2011	5532.55	5573.00	5443.55	5458.40	5447.75	5458.40	436013	1203670.01	27745700	1058350	5459.85
NIFTY	06-May-2011	26-May-2011	5474.90	5574.50	5468.00	5555.00	5547.95	5555.00	454990	1256757.59	25682300	-2063400	5551.45
NIFTY	09-May-2011	26-May-2011	5579.80	5584.10	5501.55	5558.65	5555.10	5558.65	326404	905666.77	25292450	-389850	5551.10
NIFTY	10-May-2011	26-May-2011	5557.00	5605.00	5517.00	5545.80	5549.00	5545.80	375340	1043866.05	25166350	-126100	5541.25
NIFTY	11-May-2011	26-May-2011	5553.80	5577.90	5523.50	5564.65	5562.80	5564.65	236389	656410.16	24611400	-554950	5565.05
NIFTY	12-May-2011	26-May-2011	5539.00	5577.70	5470.50	5481.35	5486.60	5481.35	385004	1062354.54	26646450	2035050	5486.15
NIFTY	13-May-2011	26-May-2011	5488.00	5622.80	5465.20	5558.90	5580.85	5558.90	576044	1602389.76	24753850	-1892600	5544.75
NIFTY	16-May-2011	26-May-2011	5534.00	5537.35	5477.35	5488.70	5490.10	5488.70	310340	854239.24	25964350	1210500	5499.00
NIFTY	17-May-2011	26-May-2011	5491.00	5521.60	5422.55	5447.40	5450.00	5447.40	376444	1033918.70	26145750	181400	5438.95
NIFTY	18-May-2011	26-May-2011	5459.85	5463.55	5397.60	5424.05	5427.00	5424.05	325424	893911.36	25987450	-158300	5420.60
NIFTY	19-May-2011	26-May-2011	5439.85	5451.80	5404.45	5424.80	5433.95	5424.80	237264	643351.15	25447050	-540400	5428.10
NIFTY	20-May-2011	26-May-2011	5440.00	5522.40	5428.80	5484.60	5479.00	5484.60	402198	1102522.32	24554650	-892400	5486.35
NIFTY	23-May-2011	26-May-2011	5439.00	5439.90	5364.75	5383.45	5382.00	5383.45	355103	969484.85	23788550	-766100	5386.55
NIFTY	24-May-2011	26-May-2011	5386.00	5417.80	5360.00	5387.85	5393.90	5387.85	374494	1009799.30	20787850	-3000700	5394.85

Table 1: A Sample of End of Day Data, taken from NSE website (www.nse-india.com)

Minute Data:

Considering the above problem of capturing intra-day fluctuations of the market, we can look at another type of data: minute data. It is the last traded price (LTP) for every derivative, recorded at the end of each minute that trading takes place. Such kind of data can be made available via a BLOOMBERG terminal, or a Thomson-Reuters screen, which is capable of recording such data live, as the market progresses.

Tick-Data:

While minute data is much more informative than end of day data when it comes to intra-day trades, nowadays, we can move even further into the details of the prices, with tick-data. Tick-data captures all the variations that take place in the prices of a derivative even to the extent of detail of intra-second variations, which themselves vary as much as 100 times in a single second, depending on the movement of trends in the market on various days. Tick data can also be made available from BLOOMBERG or Thomson-Reuters. Such data can also be, directly purchased from the NSE, or any other relevant exchange itself; which can be considered to be a bit more reliable than data given by the former, although the degree of difference between the two, if at all present, is not of great consequence.

Properties of Tick-Data include a Time-Stamp, a Financial Security Identification Code, a Last Traded Price (LTP), Last Traded Size. In some cases, information on implied volatility is also provided (as calculated by the Black-Scholes Model) (Aldridge I, 2002).

The table shows a sample of tick data of NIFTY future of date April 7, 2011.The table shows tick data capturing each trade in NIFTY futures even if it is intra second.

Date & time		Traded price	Traded volume
4/7/2011 9:15:13	TRADE	2793.15	2
4/7/2011 9:15:14	TRADE	2792.95	2
4/7/2011 9:15:14	TRADE	2793.1	1
4/7/2011 9:15:16	TRADE	2793	2
4/7/2011 9:15:21	TRADE	2793.15	1
4/7/2011 9:15:23	TRADE	2795	25
4/7/2011 9:15:25	TRADE	2795	5
4/7/2011 9:15:26	TRADE	2795	2
4/7/2011 9:15:27	TRADE	2795	2
4/7/2011 9:15:27	TRADE	2795	2
4/7/2011 9:15:29	TRADE	2795	11
4/7/2011 9:15:30	TRADE	2794.95	2
4/7/2011 9:15:31	TRADE	2794.95	1
4/7/2011 9:15:33	TRADE	2794.3	1
4/7/2011 9:15:34	TRADE	2793.15	4
4/7/2011 9:15:37	TRADE	2794	1
4/7/2011 9:15:38	TRADE	2794.65	2
4/7/2011 9:15:39	TRADE	2794	4
4/7/2011 9:15:40	TRADE	2795	9
4/7/2011 9:15:42	TRADE	2795	4
4/7/2011 9:15:43	TRADE	2795	1
4/7/2011 9:15:44	TRADE	2795	3
4/7/2011 9:15:45	TRADE	2795	2
4/7/2011 9:15:47	TRADE	2794.15	1
4/7/2011 9:15:48	TRADE	2794.4	2

Table 2: A Sample of Tick Data, as given by Bloomberg

Section 3: Merging

The above forms of data are available for each kind of exchange traded derivative. As derivatives can be hedged, it is very usual that two or more correlated derivatives are traded simultaneously, in order to hedge the risk effectively. The accurate analysis requires the merging of data of those derivatives to get their respective prices at the same instant of time.

Tick data refers to any market data which shows the price and volume of every point in increasing order of time. In order to analyze strategies we require data pertaining to all the component derivatives at that particular instant of time. In highly traded derivatives, there is more than one tick data per second. Consider derivative(X) and derivative(Y) to be a part of the strategy that is under consideration.

The first entry of derivative(X) may not occur at the same time as that of derivative(Y) i.e., the first tick of derivative(X) may be in the first quarter of the second, while the first tick of derivative(Y) may be in the last quarter of the second. (Falkenberry T.N., 2002).

This discrepancy needs to be ignored, as a greater level of accuracy in time for every tick will be required to resolve it. If tick-data time is made available at millisecond accuracy, then this problem might be resolved. Another problem is that the number of ticks per second of derivative(X) need not be the same as that of derivative(Y). So, by some merging method, the number of tick data entries needs to be made same for all the component derivatives of the trading strategy, so that ordered pairs of data are available for further analysis.

In order to resolve the above problem the following Five Merging Algorithms are suggested:

- Drop Down
- Truncation
- All Combination
- Weighted Mean
- Median

The data on 'S&P CNX Nifty' Futures and Options has been considered henceforth to analyze the merging methods.
Tables 3 and 4 represent unmerged data of two derivative securities, derivative(X) and derivative(Y), for one second, which will be used as a sample to explain the proposed merging algorithms.

Derivative (X) – (12 Ticks)

Date	Time	Price	Quantity
20090706	10:00:00 AM	4449.8	10
20090706	10:00:00 AM	4448	20
20090706	10:00:00 AM	4448.55	10
20090706	10:00:00 AM	4449.85	10
20090706	10:00:00 AM	4447.6	15
20090706	10:00:00 AM	4448	10
20090706	10:00:00 AM	4448	10
20090706	10:00:00 AM	4448	30
20090706	10:00:00 AM	4448	20
20090706	10:00:00 AM	4448	10
20090706	10:00:00 AM	4448	25
20090706	10:00:00 AM	4448	10

Table 3: Tick data for Derivative X

Derivative (Y) – (7 Ticks)

Date	Time	Price	Quantity
20090706	10:00:00 AM	193.3	10
20090706	10:00:00 AM	193.2	10
20090706	10:00:00 AM	193.2	20
20090706	10:00:00 AM	193.3	10
20090706	10:00:00 AM	193.2	10
20090706	10:00:00 AM	193.3	15
20090706	10:00:00 AM	193.3	10

Table 4: Tick data for Derivative Y

1. Drop Down

This algorithm is the most widely used one for merging. As mentioned before, it is used by most of the trading companies to merge data and also by organizations which provide tick-data. This algorithm repeats the last tick-value of the second for the series with lesser number of ticks in that second.

a) Algorithm:

- Initial Contact: Data of all series may not start from the same time, so we need to remove data till the time when all component derivatives have started being traded.
    - Convert the date and time into a number (Matlab's 'datenum' function may be used) for all derivatives, referred as time-value henceforth.
    - Now, check the first time-value (value obtained after converting date and time) of all the series and the largest one is the 'first contact'.
    - Skip entries in all tick data of Derivatives till the time-value becomes equal to that of 'first contact'.
- Time Loop: We will have to run a loop for the incrementing the time one second per iteration.
    - We check for the time-value of one second (=1.1574e-5 by Matlab's 'datenum' function).
    - Now, run a 'for' loop for time starting with the 'first contact' incrementing the loop-time (which starts with the 'first contact') by the time-value of a second for every loop.
    - Calculate the maximum number of seconds for which trade occurs and run the loop that many times (e.g. from 9:15:00 AM to 3:30:00 PM, we have 22501 seconds, therefore the 'for' loop needs to be run 22501 times).
- Second Loop: We now need to make combinations for values that have the same time of trade.
    - Take the difference between the loop time and the time-value of the derivatives at their current pointer locations.
    - Run a 'while' loop nested within the previous 'for' loop, which exits only if all the differences are non-zero (i.e., even if one difference is non-zero, the loop continues).
    - Copy the current time-value of the loop in the first column of a new matrix (merged_file).
    - Run 'if' loop within the 'while' loop for all the derivatives individually.
        - If the difference is zero, copy the price of the derivative into the matrix (merged_file) and move the pointer to the next entry in the tick data table of that derivative.

- If the difference is not zero, then copy the previous value into the matrix (merged_file)

b) Output: The matrix (merged_file) is the required merged file, with same number of tick data entries per second for all derivatives.

Advantage:

- This model captures the volatility of the market.
- It accounts for all the individual trades that have taken place in the market.

c) Disadvantage:

- Lower frequency data gets dragged, excess dragging may change the nature of data
- Excess dragging leads to misrepresentation of data.

d) Output File:

Date&Time	price(X)	price(Y)
20090706 10:00:00 AM	4449.8	193.3
20090706 10:00:00 AM	4448	193.2
20090706 10:00:00 AM	4448.55	193.2
20090706 10:00:00 AM	4449.85	193.3
20090706 10:00:00 AM	4447.6	193.2
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3

Table 5: Results of Data merged by the Drop Down Algorithm

2. Truncation

This algorithm skips the excess ticks of the second for the series with larger number of ticks in that second.

a. Algorithm:

- Initial Output File: Creating a file into which the merge data will be dumped.
  - Convert the date and time into a number (Matlab's 'datenum' function may be used) for all derivatives, referred as time-value henceforth.
  - Copy Derivative (derivative(Y)) with lesser tick data frequency (approx.) into the initial output file.

- Merging: The tick data values of the other derivative will be entered in the initial output file.
  - Run a 'for' loop for the number of data in derivative(Y).
  - Run a nested 'for' loop for the number of data of derivative(X)
  - Check if the time-value of derivative(Y) is same as that of derivative(X).
    1. If the time-value is same, then copy the price of derivative(X) in the initial output array corresponding to the derivative(Y) price at that time-value and increment the pointer to the next tick-value of derivative(X) and make it the starting point for the next 'for' loop for derivative(X).
    2. If the time-value of derivative(Y) is greater, then let the 'for' loop go to the next iteration and then check the time-values.
    3. If the time-value of derivative(X) is greater, then put a zero in the initial output array corresponding to the derivative(Y) price at that time-value and exit the 'for' loop for derivative(X) .

- Final Output File: Remove zeros in price column
  - Check for zeros in prices of the derivatives (find() in Matlab can be used), and remove those entries (an index with entries where derivative prices are non-zero can be created and only those values can be copied to the final output file).
  - The 'Final Output File' is the required merged file, with same number of tick data entries per second for all derivatives

b. Advantages:

- No dragging effect is seen.
- The trend is captured by less data.

c. Disadvantages:

- A major chunk of data is removed if the difference in number of tick data per second, between different derivatives is high.
- Jumps are introduced in data whose path cannot be accounted for.

d. Output File:

Date &Time	price(X)	price(Y)
20090706 10:00:00 AM	4449.8	193.3
20090706 10:00:00 AM	4448	193.2
20090706 10:00:00 AM	4448.55	193.2
20090706 10:00:00 AM	4449.85	193.3
20090706 10:00:00 AM	4447.6	193.2
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448	193.3

Table 6: Results of Data merged by Truncation Algorithm

3. All Combinations

This algorithm makes all possible combinations within a second between ticks of all the series.

i. Algorithm

- Removing Repetitions: Remove repetitions within a second for all the series
  - Convert the date and time into a number (Matlab's 'datenum' function may be used) for all derivatives.
  - Run a loop for incrementing the time by 1 second per iteration
  - Check for repeated prices within a second and remove them.
- Forming Combinations: All possible combinations are formed.
  - Run a loop for incrementing the time by 1 second per iteration with starting point as the point of initial contact (refer drop down algorithm for the method to find initial contact).
  - Copy the time-value into the first column of output array (merged_file)
  - Map each entry of derivative(X) with all entries of derivative(Y) within the second. (extend it to all the component derivatives) and copy them to output array (merged_file)
- Output: The array (merged_file) is the required merged file, with all possible combinations of ticks within a second.

ii. Advantages:

- All possible combinations of trades occurring within the second are made, so some of them will capture the actual pairs.

iii. Disadvantages:

- This algorithm is repeating trends or making step functions.
- Creation of unnecessary combinations, excess of which may change the nature of data.

iv. Output File:

Date &Time	price(X)	price(Y)
20090706 10:00:00 AM	4449.8	193.3
20090706 10:00:00 AM	4448	193.3
20090706 10:00:00 AM	4448.55	193.3
20090706 10:00:00 AM	4449.85	193.3
20090706 10:00:00 AM	4447.6	193.3
20090706 10:00:00 AM	4449.8	193.2
20090706 10:00:00 AM	4448	193.2
20090706 10:00:00 AM	4448.55	193.2
20090706 10:00:00 AM	4449.85	193.2
20090706 10:00:00 AM	4447.6	193.2

Table 7: Results of merging Data by the All-Combinations Algorithm

Derivative(X) – After removing duplicates:

Date &Time	price(X)
20090706 10:00:00 AM	4449.8
20090706 10:00:00 AM	4448
20090706 10:00:00 AM	4448.55
20090706 10:00:00 AM	4449.85
20090706 10:00:00 AM	4447.6

Table 8: Derivative X, without Repetitions

Derivative(Y) – After removing duplicates:

Date &Time	price(X)
20090706 10:00:00 AM	193.3
20090706 10:00:00 AM	193.2

Table 9: Derivative Y, without Repetitions

4. Weighted Mean

This algorithm considers the weighted mean of prices of every second (traded quantity is the weight) of all the derivatives and merges them by truncation/ dropdown algorithm.

I. Algorithm:

- Calculate the Weighted Mean of the tick data for every second (traded quantity is the weight) and copy it to an output file.
- There remains either one or zero ticks for each second in the output file.
- Merge the output files of all the derivatives using Truncation/ Drop down to get the Final Output file (In this paper, Truncation is used for merging Output file).

II. Advantages:

- The trend is captured in less data
- The variations within the entire second are considered

III. Disadvantages:

- It does not account for the number of ticks that are present within a second (i.e., equal value to a second with 1 tick and a second with 100 ticks).

IV. Output File:

Date &Time	price(X)	price(Y)
20090706 10:00:00 AM	**4448.2**	**193.25**

Table 10: Results of data merged by the Weighted Mean Algorithm

5. Median

This algorithm considers the median of prices of every second for all the derivatives and merges them by truncation/ dropdown algorithm

I. Algorithm:

- Extract the median for every second of the tick data for all the derivatives and copy it into the output file
- There remains either one or zero ticks for each second in the output file.
- Merge the newly generated files (median files) using truncation algorithm

II. Advantages:

- The trend is captured in less data

III. Disadvantages:

- Minute variations of data within a second are not captured.
- In a high volatility market, the algorithm may not be able to capture the trend

IV. Output file:

Date &Time	price(X)	price(Y)
20090706 10:00:00 AM	**4448**	**193.3**

Table 11: Results of Data merged by the Median Algorithm

4. Analysis

Before we move on to the results, we must first discuss: what is the need for the data to be of a specific type? More specifically, since while merging, we try to retain the nature of the data, we must first establish what is meant by 'nature' of the data. We are describing the nature of the data using three standards, viz. Stationarity, Normality, and Randomness of the data.

Stationarity of data implies that the mean and variance of the data is not varying with respect to time. This means, if we select any segment of data for any duration, the mean and variance of the data will be the same if the data is stationary. It becomes very easy to analyze data if it is stationary, rather than non-stationary.

Normality of the data basically gives us a hint of the distribution of the data. If the data is found out to be normal, it can be of a great consequence, as then the nature of the data will be very well known, and definite confidence intervals for the variation of the data ranges can be made out with certainty.

Randomness of the data refers to the property of the data whereby the forthcoming or latter values of the data are not dependent upon the previous or former values of the respective data. This property was obviously difficult to establish, as data in stock markets will not definitely be random in nature. Randomness of the data is important to establish normality of data. Also, it is important to compare the randomness of pre-merged and post-merged data, in order to find out if the nature of data has been retained or not.

Current Context

Before conducting any stastistical analysis on any data series, the necessary and sufficient condition is to be confirm about the nature of data.In order to elucidate the statistical differences between the three forms of data ,certain tests on the nature of data were conducted.

The following tests were conducted on all 3 forms of data:

- **ADF/KPSS test:** for checking the stationarity of data series.
  h = adftest(*y*) assesses the null hypothesis of a unit root in a univariate time series *y*.
- **Runs test:**
  h = runstest(x) performs a runs test on the sequence of observations in the vector x. This is a test of the null hypothesis that the values in x come in random order, against the alternative that they do not. The test is based on the number of runs of consecutive values above or below the mean of x.

- **Jarque bera test**:
  h = jbtest(x) performs a Jarque-Bera test of the null hypothesis that the sample in vector x comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.

Nifty F&O	Stationarity ADF/KPSS Test	Randomness Runs Test	Normality JB Test
LTP (last Traded Price) of the Day			
Price	Doesn't Exist	Doesn't Exist	Doesn't Exist
First Difference	Exists	Exists	Exists
Return	Exists	Exists	Exists
Minute Data			
Price	Doesn't Exist	Doesn't Exist	Doesn't Exist
First Difference	Exists	Exists	Doesn't Exist
Return	Exists	Exists	Doesn't Exist
Tick Data			
Price	Doesn't Exist	Doesn't Exist	Doesn't Exist
First Difference	Exists	Doesn't Exist	Doesn't Exist
Return	Exists	Doesn't Exist	Doesn't Exist

Table 12:Results of various statistical tests involving nature of data

Price of derivative at time 't' : $\mathbf{p_t}$

First Difference of Price : $\mathbf{p_t - p_{t-1}}$

Return of Price : $\mathbf{(p_t - p_{t-1})/p_t}$,

All the test were conducted on the prices of Nifty Futures,1st difference($\mathbf{(p_t - p_{t-1})}$, where p_t dentes price at time t)) of the prices and their return $\mathbf{((p_t - p_{t-1})/p_t}$, where p_t dentes price at time t) obtained as end of the day data, minute data and tick data.

It is evident from table 12 that while prices of Nifty futures are not random,stationary or normal in all the data forms.But while 1st difference and return of end of the day prices are all staionary, normal as well as random.Minute data 1st differnce and returns are random and stationary.Tick data 1st differnce and returns are only stationary. \

Minute Data Analysis

Minute Data			Min		
21-Jan-11			27-Dec-10		
Trend - Neutral Vol - Low			Trend - Down Vol - Low		
Man U Whitney Test			Man U Whitney Test		
Strike	First Dif	Return	Strike	First Dif	Return
Future	0.4084	0.3937	Future	0.2934	0.3042
5600C	0.1753	0.1822	5900C	0.3519	0.3581
5800C	0.7002	0.7216	6100C	0.033	0.0481
5600P	0.0014	0.0011	5900P	0.5016	0.4934
5800P	0.4746	0.466	6100P	0.6228	0.5684

Table 13: Results of Mann-U-Whitney test on minute data

In statistics, the **Mann–Whitney *U*** test is a non-parametric statistical hypothesis test for determining whether two independent samples of observations have equally large values. It is one of the most well-known non-parametric significance tests. It performs a two-sided rank sum test of the null hypothesis that data in the vectors x and y are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians. x and y can have different lengths. The *p* value of the test is returned in p**. P value below .05 indicates rejection of null with 95% level of confidence.**

As none of the three forms of minute data is normally distributed (distribution is not certain), thus a non parametric test needs to be conducted.

The Mann U Whitney test was conducted on the 1st differences and return of Nifty future prices as well as Call and Put prices of 2 different strikes on 2 different days taking minute data series of specific derivative as one series and tick data of derivative as another series. . The days and strikes chosen were completely random. The trend in the market on those specific days is mentioned in the header of table 13. As seen from table 13, the minute data fails the test on both the days for two different NIFTY options indicating that minute data and tick data series for the specific derivative are not statistically same. As we know that tick data lists all the trades in a given derivative .Also the above mentioned test compares distribution of data about its median, it is clear that minute data brings about a change in the distribution of tick data about its median not preserving its nature. Thus we can say that:

- Minute data does not capture all the fluctuations of the market.
- Minute data does not always retain nature of actual traded data series

Comparison of merging algorithms

		Number of Strikes in p-value range of Man U Whitney Test:								
		Red	%	Yellov	%	Blue	%	Green	%	Total
Weighted Mean										
WM	Rank:2	2	6.25	10	31.25	11	34.38	9	28.13	32
Truncation										
T	Rank:1	1	3.125	1	3.125	10	31.25	20	62.5	32
All Combination										
AC	Rank:4	6	20	0	0	11	36.67	13	43.33	30
Drop Down										
DD	Rank:5	8	28.57	2	7.143	2	7.143	16	57.14	28
Median										
M	Rank:3	5	20.83	4	16.67	8	33.33	7	29.17	24
Colour codes & Ranges:			0-0.05		0.05-0.15		0.15-0.5		0.5-1	No Data

Table 14:Comparison of merging algorithms based on Mann U Whitney test

Checking for more efficient merging algorithm is being done based on the nature of pre-merged and post-merged data.Pre merged and post merged data of specific deriavtive have been taken as the two different series on which **Mann U whitney test** was conducted. As mentioned above, P value less than .05 rejects the null. Colour coding of the P values of the test is done , thus while all other shows acceptance ,red colour shows rejection.

Four random days were chosen and tests were conducted on 1st differnce of the prices and returns of the NIFTY future,Call and put options of 2 different strikes. Merging was done using all the 5 algorithms and number of times the results of test falling in different P values region shown by colour codes for each algorithm is listed in the table.

For the results of Mann U Whitney test to be positive, it is required that the nature of distribution about the median of raw data is the same as that of merged data. From table 14, it is found that 'Truncation' and 'Weighted Mean' techniques yield a better result than the other algorithms.Details of the 'p-values' obtained are given in Appendix-A.

Comparison of merging algorithms using Random Data

		Behaviour of p-values of Man U Whitney Test (Number of values in the Range):								
		Red	%	Faint Green	%	Light Green	%	Green	%	Total
Weighted Mean										
WM	Rank:	0	0	0	0	0	0	8	100	8
Truncation										
T	Rank:	0	0	0	0	1	12.5	7	87.5	8
All Combination		**Randomness** of Original Sample is not preserved, **Normality** of option(4) is also not preserved.								
AC	Rank:	0	0	0	0	1	12.5	7	87.5	8
Drop Down		**Randomness** of Original Sample is not preserved, **Normality** of option(4) is also not preserved.								
DD	Rank:	0	0	1	12.5	1	12.5	6	75	8
Median		**Randomness** of Original Sample is not preserved, **Non-Normality** of option(1) is also not preserved.								
M	Rank:	3	37.5	3	37.5		0	2	25	8
			Reject				Accept			
Colour codes & Ranges:										
		Red	0-0.05	Faint Green	0.05-0.15	Light Green	0.15-0.5	Green	0.5-1	

Table 15: Comparison of merging algorithms using Random Data, based on Mann U Whitney test. (Compare results with table 14)

Random numbers are created and they are assumed to be the prices of component derivatives and then the nature of distribution about median for pre-merge and post-merge data is checked. It is found that 'All Combination', 'Drop Down' and 'Median' algorithms are not able to preserve the random nature of the prices in some cases.

Also, 'Option 4' that was considered had a normal distribution', but 'All Combination', 'Drop Down' algorithms did not retain the normal nature after merging. 'Option 1' that was considered did not have a normal distribution', but 'Median' algorithm made it a normal distribution after merging. Details of the 'p-values' obtained are given in Appendix-B.

This observation supports the previous result that 'Truncation' and.'Weighted Mean' algorithms give better results.

Regression Results

- After running regression various models using tick data, residuals were found to be stationary but not normally distributed.
- Even in the case of minute data residuals were stationary but not normally distributed.
- This phenomenon is not specific to Indian Markets but is found even in the case of data corresponding to foreign indices.

❖ The Regression Equation used cannot be divulged due to confidentiality.

Conclusion

Tick data empowers us to analyze the fluctuations of prices of derivatives & strategies within a second. We need a method to merge data as we need prices of different derivatives at the same time instant. 'Truncation' and 'Weighted Mean' algorithms were found the most efficient merging algorithms discussed so far. But, for regression analysis, tick data may not yield accurate results due to non-normal nature of residues.

References

Shah, A, and Thomas, S "Securities Markets—Towards Greater Efficiency." In *India Development Report*, edited by K. Parikh. UK: Oxford UniversityPress, 2002.

Falkenberry, T. N., "High Frequency Data Filtering", as seen on www.tickdata.com, 2002

Aldridge I, "High Frequency Trading", John Wiley and Sons, 2010

Websites:

- http://www.nseindia.com
- http://www.finance.lab.iimcal.ac.in/abt.asp
- http://www.universityfinancelab.com/
- www.mathworks.com
- www.world-exchanges.org/

Appendix-A

Future with Call Merge	21-Jan-11 Trend - Neutral Vol - Low			27-Dec-10 Trend - Down Vol - Low			16-Jun-09 Trend - Up Vol - High			5/9/2011 SBIN Trend - Neutral Vol - Me		
	Man U Whitney Test			Man U Whitney Test			Man U Whitney Test			Man U Whitney Test		
	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return
Weighted Mean	5600	0.4999	0.4876	5900	0.2701	0.2699	4300	0.0433	0.0435	2600	0.4553	0.4658
WM	5800	0.6281	0.6101	6100	0.2562	0.2634	4600	0.2062	0.2057	2700	0.1045	0.1023
Truncation	5600	0.7203	0.674	5900	0.1527	0.1519	4300	0.0408	0.0391	2600	0.7761	0.7859
T	5800	0.3345	0.2854	6100	0.362	0.3515	4600	0.3017	0.2917	2700	0.3333	0.3265
All Combination	5600	0.0244	0.0356	5900	0.0151	0.0182	4300	0.633	0.6516	2600	0.1449	0.1491
AC	5800	0.0088	0.0049	6100	0.0427	0.0401	4600	0.8162	0.8257	2700	0.0013	0.0013
Drop Down	5600	0.0069	0.0067	5900			4300	0.9983	0.9983	2600	0.9985	0.9964
DD	5800	0.0092	0.009	6100	0.0348	0.0348	4600	0.949	0.949	2700	0.852	0.8518
Median	5600	0.2593	0.2248	5900	0.1121	0.1042	4300	0.0281	0.0305			
M	5800	0.056	0.0721	6100	0.0377	0.0308	4600	0.2444	0.253			
Winner:	**WM**			**WM**			**DD**			**DD**		
Failure:	**AC**	**DD**		**AC**	**DD**	**M**	**WM**	**T**	**M**	**AC**		

Colour codes & Ranges: 0-0.05 | 0.05-0.15 | 0.15-0.5 | 0.5-1 | No Data

Call Merge	21-Jan-11 Trend - Neutral Vol - Low			27-Dec-10 Trend - Down Vol - Low			16-Jun-09 Trend - Up Vol - High			5/9/2011 SBIN Trend - Neutral Vol - Me		
	Man U Whitney Test			Man U Whitney Test			Man U Whitney Test			Man U Whitney Test		
	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return
Weighted Mean	5600	0.5734	0.5437	5900	0.0511	0.0551	4300	0.768	0.7635	2600	0.2714	0.2902
WM	5800	0.1001	0.0977	6100	0.0709	0.0714	4600	0.139	0.1409	2700	0.1867	0.2097
Truncation	5600	0.949	0.9619	5900	0.7531	0.7453	4300	0.9	0.8982	2600	0.845	0.8216
T	5800	0.6113	0.6371	6100	0.4113	0.3675	4600	0.9081	0.9213	2700	0.5172	0.5334
All Combination	5600	0.4345	0.4089	5900	0.8641	0.8629	4300	0.8228	0.8216	2600	0.3145	0.3196
AC	5800	0.3093	0.3086	6100	0.1062	0.3393	4600	0.8406	0.8252	2700		
Drop Down	5600	0.7924	0.7924	5900			4300	0.5758	0.5758	2600	0.084	0.084
DD	5800	0.8712	0.8712	6100	0.9707	0.9707	4600	0.8327	0.8327	2700	0.067	0.0667
Median	5600	0.9161	0.9692	5900	0.2564	0.2603	4300	0.8805	0.8809			
M	5800	0.7451	0.7213	6100	0.0911	0.1263	4600	0.3156	0.3174			
Winner:	**M**			**DD**			**T**			**T**		
Failure:	**x**			**x**			**x**			**x**		

Colour codes & Ranges: 0-0.05 | 0.05-0.15 | 0.15-0.5 | 0.5-1 | No Data

Put Merge	21-Jan-11 Trend - Neutral Vol - Low			27-Dec-10 Trend - Down Vol - Low			16-Jun-09 Trend - Up Vol - High			5/9/2011 SBIN Trend - Neutral Vol - Me		
	Man U Whitney Test			Man U Whitney Test			Man U Whitney Test			Man U Whitney Test		
	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return
Weighted Mean	5600	0.4906	0.4965	5900	0.3475	0.3604	4300	0.5251	0.5255	2600	0.7758	0.7765
WM	5800	0.1712	0.1524	6100	0.0507	0.0615	4600	0.4286	0.4354	2700	0.7763	0.7756
Truncation	5600	0.7602	0.7669	5900	0.7654	0.7562	4300	0.9768	0.9734	2600	0.7696	0.745
T	5800	0.9082	0.9062	6100	0.5435	0.5539	4600	0.9912	0.9953	2700	0.9079	0.9187
All Combination	5600	0.3505	3067	5900	0.9134	0.9246	4300	0.4284	0.3371	2600	0.7817	0.8134
AC	5800	0.6228	0.6024	6100	0.3928	0.5872	4600	0.2298	0.2182	2700		
Drop Down	5600	0.5839	0.5841	5900			4300	0.8838	0.8838	2600	0.2755	0.2756
DD	5800	0.4787	0.4787	6100	0.8306	0.8306	4600	0.0419	0.0419	2700	0	0
Median	5600	0.4661	0.5469	5900	0.0559	0.064	4300	0.4025	0.4089			
M	5800	0.9448	0.9789	6100	0.0391	0.0482	4600	0.5867	0.5795			
Winner:	**T**			**DD**			**T**			**T**		
Failure:	**x**			**M**			**DD**			**DD**		

Colour codes & Ranges: 0-0.05 | 0.05-0.15 | 0.15-0.5 | 0.5-1 | No Data

Future with Put Merge	21-Jan-11 Trend - Neutral Vol - Low			27-Dec-10 Trend - Down Vol - Low			16-Jun-09 Trend - Up Vol - High			5/9/2011 SBIN Trend - Neutral Vol - Me		
	Man U Whitney Test			Man U Whitney Test			Man U Whitney Test			Man U Whitney Test		
	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return	Strike	First Diff	Return
Weighted Mean	5600	0.5929	0.5743	5900	0.1625	0.1528	4300	0.0057	0.006	2600	0.2517	0.2478
WM	5800	0.4582	0.4714	6100	0.0714	0.0665	4600	0.1135	0.1138	2700	0.079	0.0783
Truncation	5600	0.5839	0.5217	5900	0.3761	0.3745	4300	0.6574	0.662	2600	0.1068	0.1027
T	5800	0.688	0.6424	6100	0.2265	0.2178	4600	0.2864	0.2777	2700	0.3333	0.3265
All Combination	5600	0.8489	0.6918	5900	0.2935	0.2935	4300	0.633	0.6516	2600	0.1449	0.1491
AC	5800	0.6583	0.5022	6100	0.2132	0.2132	4600	0.8162	0.8257	2700	0.0013	0.0013
Drop Down	5600	0.0069	0.0067	5900			4300	0.9983	0.9983	2600	0.9985	0.9964
DD	5800	0.0092	0.009	6100	0.0348	0.0348	4600	0.949	0.949	2700	0.852	0.8518
Median	5600	0.1971	0.2415	5900	0.1523	0.1326	4300	0.0121	0.0126			
M	5800	0.8823	0.814	6100	0.0329	0.0279	4600	0.2386	0.2397			
Winner:	**AC**			**T**			**DD**			**DD**		
Failure:	**DD**			**DD**	**M**		**T**	**M**		**AC**		

Appendix-B

Future		Random				
		Man U Whitney Test				
		Option Size:	very low(4)	low(3)	med(2)	high(1)
Weighted Mean						
WM			0.8647	0.5795	0.6886	0.8561
Truncation						
T			0.526	0.7841	0.8505	0.6748
All Combination						
AC			Randomness of Original Sample is not preserved			
			0.9822	0.8713	0.7381	0.9326
Drop Down						
DD			Randomness of Original Sample is not preserved			
			0.992	0.9954	0.7865	0.9485
Median						
M			Randomness of Original Sample is not preserved			
			0.0763	0.0738	0.0473	0.0766
Colour codes & Ranges:			Reject			Accept

Option		Random				
		Man U Whitney Test				
		Option Size:	very low(4)	low(3)	med(2)	high(1)
Weighted Mean						
WM			0.7208	0.3243	0.5352	0.9079
Truncation						
T			0.8504	0.5512	0.5795	0.7376
All Combination						
AC			Randomness of Original Sample is not preserved			
			0.4991	0.9977	0.8659	0.716
Drop Down						
DD			Randomness of Original Sample is not preserved			
			0.5253	0.3689	0.1475	0.7879
Median						
M			Randomness of Original Sample is not preserved			
			0.7988	0.0201	0.7354	0.0289
Colour codes & Ranges:			Reject			Accept