Forecasting Daily Market Direction: A Data Mining Case Using the Nasdaq

Mary E. Malliaris
Loyola University Chicago

Steven G. Malliaris
Massachusetts Institute of Technology

Abstract. This study looks for evidence that the US market is influenced by other international markets in an amount significant enough to give us daily clues about the direction the market will move. The target US market chosen was the Nasdaq. The international markets selected for inputs into the model include Australia, Japan, and Hong Kong. This study uses one of the most popular data mining techniques, the decision tree. The decision tree forecasts work significantly better on days when the Nasdaq was Up than it did on Down days. The directional forecasts on Up days are correct about two-thirds of the time.

1. Introduction

Studies on international market co-movement have shown differing results over the years. Hilliard (1979) studied ten world exchanges, including Tokyo, Sydney, and New York, and concluded that inter-continental prices did not seem to be closely related. Darbar and Deb (1994) studied four markets, Canada, Japan, UK, and US, for 1989 through 1992. Their conclusion was that prices are not cointegrated. Specifically, they state that "the Japanese and U.S.A. stock markets have episodes of significant correlation, but have zero permanent correlation." Dwyer (1998) looked at monthly data for 31 years from stock indices for the US, UK, Japan and Germany and concluded that there was no evidence that the levels of the indices are related.

In contrast, Fischer and Palasvirta (1990) studied twenty-three market indices from around the world and discovered a statistically significant level of interdependence, with the US index prices leading other countries. Ribeiro and Veronesi (2002) found that correlations of international market returns increase during bad times. Brooks, Forbes, and Mody (2003) found that "financial comovement has increased in the 1990s", though "real comovements are neither especially strong nor on an obviously upward trend". For an general list of studies on international linkages over the last twenty years, including lead-lag relationships, see Capelle and Raymond (2002).

Over time, however, many studies are finding two results. The first is that correlations between international markets fluctuate significantly over the years, and the second is that the US indices lead other markets.

This study uses data mining to look for evidence that the US market is influenced by other international markets in an amount significant enough to give us daily clues about the direction the market will move. In other words, we look for evidence that the US market is led by other markets to some degree. Thus the target market for forecasting in this study is a US market. The target US market chosen was the Nasdaq. The international markets selected for inputs into the model include the markets of Australia, Japan, and Hong Kong. Each of these markets has opened and closed on a given day before the Nasdaq has opened, and the Nasdaq is closed before any of these markets open for the next day, thus restricting the analysis to net change in price and eliminating possible extraneous predictive effects that could be caused by comovement amongst simultaneously open markets. This approach also allows a directional forecast to be made before the Nasdaq has begun trading on a given day, eliminating effects of news happening while this market is open. This study also limits the scope to a daily effect from these three Eastern markets. No lags of information are included. The focus is restricted simply to daily information from three markets that have opened and closed for the day while the Nasdaq was closed for the night.

2. Data

Daily closing price data was collected from Yahoo!Finance for each of four markets: Nasdaq Composite (IXIC), Australia All Ordinaries (AORD), Hong Kong Hang Seng (HSI), and Japan Nikkei 225 (N225). The data downloaded began in January, 1998 and continued through December, 2002. Daily data from Australia, Japan, and Hong Kong was used as inputs to forecast the Nasdaq direction on the same day.

This daily closing price data was used to form calculated fields for each market consisting of percent change in closing price (market close today minus market close yesterday divided by market close yesterday), direction of market change (Down if today minus yesterday was less than zero, Up otherwise), and the number of nights a market was closed locally (today's date minus date market was last open). These fields were labeled as chgC, Dir, and Days, respectively. These three fields were calculated for each of the input markets and were labeled for Australia, Japan and Hong Kong, respectively, as AchgC, ADir, ADays, JchgC, JDir, JDays, HKchgC, HKDir, and HKDays. Days and Dir were also calculated for the Nasdaq and labeled as NDays, and NDir.

The inclusion of a variable tracking the number of days (ADays, JDays, HKDays, and NDays) since the particular market was last open is based on the findings of Malliaris and Salchenberger (2002). Any artificial intelligence system must have enough information to be able to mimic the decisions that might be made by an actual human in the same situation. The amount of time elapsed since the last trading day is a strong psychological influence, since the passage of time permits information to settle into one's consciousness, while simultaneously raising anxiety about trading opportunities that cannot be implemented while the market remains closed—effects which are not otherwise discernable from price movements.

The variables Dir and ChgC both track market movement, but in different ways. ChgC records the exact percentage of market change and is a real number. Dir is a symbolic field that indicates only gross movement, that is, was the market Up or Down. By including both of these measurements on each market, we can isolate the affects of both simple direction and magnitude.

After each calculated field was computed in Microsoft Excel, the data from each individual market was exported to Microsoft Access as a table. One table was constructed for each market. Date was used as the key in each table. A query was then run using all data from each table. This query then selected only those rows from each table for which a corresponding date existed in each table. Thus, if any one of the four markets was closed on a given day, that day did not come up in the query.

The query results were then exported to Excel where it was divided up into year-long segments, 1998 through 2002. Each individual year was saved as a text file for importing into SPSS.

In addition to the variables from Australia, Japan and Hong Kong already mentioned as inputs, NDays was also used as an input, and NDir was the target or forecasted variable. A sample row of values for one day is shown in Table 1.

Table 1. *Variable Values In One Example Row Of Data.*

Not Used	Input	Input	Input	Input	Input	Input	Input	Input	Input	Input	Output
Date	ADir	AchgC	ADays	JDir	JchgC	JDays	HKDir	HKchgC	HKDays	NDays	NDir
4/7/1998	Down	-.029	1	Up	1.738	1	Down	-0.03	4	1	Down

Since prior studies have found that the correlations between market changes are not significant, prior to construction of the decision trees, correlation coefficients were calculated on the change in closing price for all markets for each year. If the coefficients are high, we could expect the decision tree to gain a lot of insights from the data. These coefficients are shown in Table 2. As can be seen, the correlations between the Nasdaq and the selected Eastern markets are neither high nor stable. They would appear to indicate that not much information can be gained from the closing prices of these markets concerning the direction the Nasdaq moved that day. They show variation over years within each of the markets and variation from year to year. The market most highly correlated in one year may be the lowest in the following year. For example, in 1998 HSI had the highest correlation with the Nasdaq, while in 1999, it had the lowest.

Table 2. *Correlations Between Nasdaq And Other Market Daily Closing Prices Changes.*

Year	AORD	J225	HSI
1998	.14	.16	.25
1999	.13	.05	.01
2000	.08	.03	.14
2001	.19	.18	.21
2002	.06	.13	.09

These numbers support the findings mentioned in the Introduction that correlations vary widely over the years and are not stable within markets.

Following the creation of the yearly data sets, each set was read into the SPSS data mining package Clementine and attached to a type node to specify the level of data for each field. All Dir fields were symbolic, all ChgC fields were real, and all Days fields were integer. The data was now ready to be analyzed.

3. Methodology

This study uses one of the most popular data mining techniques, the decision tree. Rather than specifying the variables to be used before running a decision tree algorithm, this data mining techniques derives the tree from the data it analyzes (Hand, 2001). It uses only the variables that influence the path development. A decision tree is driven by one category-type output variable. Before the tree begins, all rows of data are in one group. When the tree algorithm ends, each final group will have rows containing only one value of the output variable. At each step of the tree development, all unused input variables are checked by the algorithm and the one selected is that whose values can be used to divide the data into groups with minimal variation on the output variable within groups and maximum variation between groups (Groth, 1998). Thus, the tree is built from the root to the end nodes by recursively splitting the rows at a node into two or more parts, each leading to another node (Berry and Linoff, 2000). Each node is more similar in values of the output variable than the previous node.

Each group becomes a new node in the tree. The process repeats, with a new variable picked at each node until either all variables have been used in the path, or no variable remaining contributes to the reduction of variation within the group. A rule is generated for each path from the root of the tree to each end node. These rules give insight into the thinking followed by a trader since they specify the order in which a decision is made and the variables that contribute to that decision.

The package used in this study was Clementine by SPSS. Clementine uses a decision tree algorithm known as C5.0. The C5.0 algorithm only includes variables in each path that impact the decision. It also generates a rule for each path that explicitly displays the reasoning used for the splits in the path. For details of the process, see the Clementine User Guide (1998).

When the Clementine training of the tree is completed, Clementine

generates a trained model. This model contains all the final rules that were uncovered during the construction of the tree. New data not used for training can then be run through the trained model to generate a forecasted value for the output variable. These values can then be compared to the actual value that occurred on that day to see how well the model holds up on new data.

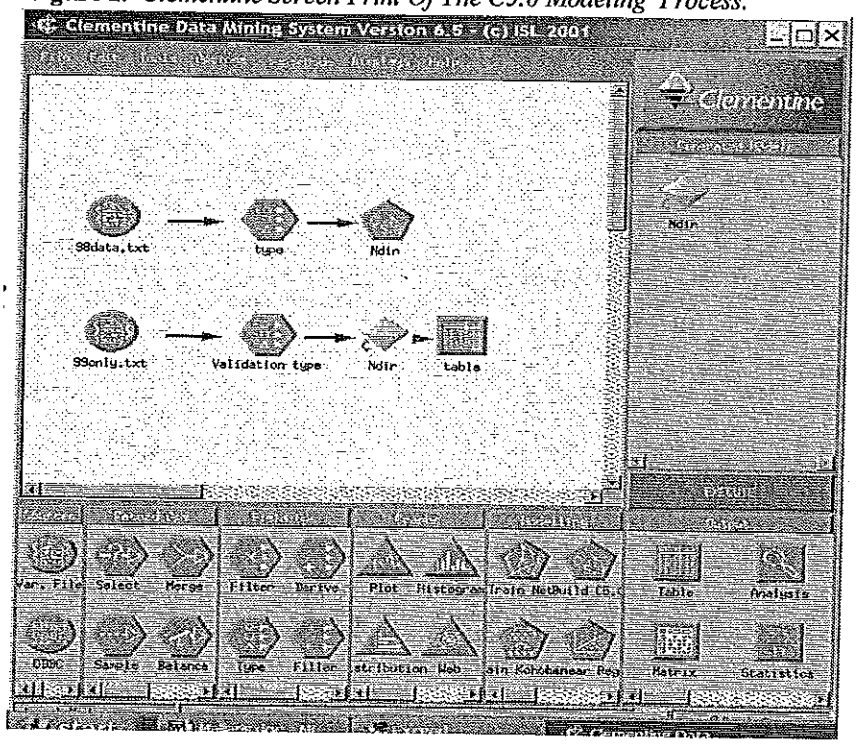
The output variable in this study was NDir, the direction of the Nasdaq on a single day. It was category-type data with two possible values, Up or Down. The input variables included direction, change in closing, and number of nights the market was closed for each of the three Eastern markets, Australia, Japan, and Hong Kong. Also used as an input was the number of nights the Nasdaq was closed. The decision tree can therefore be said to approximate the initial assumptions of a day-trader, in effect considering only that information which would have become available overnight and using it to make an overall prediction of the day's movement.

One decision tree was developed for each year of input data, that is, years 1998 through 2001, for a total of four trees. Each row of input data over all years contained only information from one day. Each of the four trained decision tree models generated a set of rules. Data from the year immediately following was run through the appropriate trained model to generate a set of daily direction forecasts for the Nasdaq. That is, the model trained on 1998 data was used to generate the forecasts for all days in 1999, the model trained on 1999 data was used to forecast all days in 2000, and so on. The output for each year was divided into days where the Nasdaq had actually been Up and those days where the actual Nasdaq movement was Down.

Figure 1 contains a screen print of one of the Clementine models. Two data nodes are shown. The upper data node is used to read in the training set, the lower data node reads in the validation set (ie, the following year). Each data node is connected to a type node where the kind of data and the inputs and output are specified. For training, the type node then connected to a C5.0 decision tree node. After training, a generated model node was created. For validation, the generated model node was connected to the validation set type node. This was then attached to a table where the results of the forecasts could be displayed. This process was carried out for each year of input data.

The forecasts were then compared to the actual directions that had occurred. Because market relationships change over time, a row of data from the validation set may satisfy each condition in the trained rule-set but may have a different value for NDir than the one the rule would predict. The number of times each forecast of direction agreed and disagreed with the directions that the Nasdaq had moved were tallied. Generated rule sets were inspected.

Figure 1. *Clementine Screen Print Of The C5.0 Modeling Process.*



4. Results

Rules sets were generated as an output for each of the four trained models. These rule sets break the entire set of data rows into groups based on the input variable values. Each group of rows represents one branch in the decision tree and all rows in that branch have the same final value on the forecasted variable, NDir, in the training set. The rules are easily interpretable. An average of twenty rules were generated for each year. Examples of one Up and one Down rule for each year are shown below in Table 3.

Looking, for example, at the Up rule for 1998, we see that three variables were used in constructing this branch of the decision tree. At the first split, a group was formed whose values on the percent change in closing price for the market in Japan was between -1.678 and -0.963. This group was then split on the number of nights the Australian market had been closed. The only rows included were ones where the Australian market had been closed only overnight, that is, number of nights closed equaled one. The final split was based on the direction the Australian market had moved. Only rows were included where the Australian market movement that day had been Up. The statement of the rule is then followed by two numbers. The first gives the number of times the "If" conditions were found in the

training set. In other words, it tells us how many rows from the training set were used to generate each branch of the decision tree. The second number gives the probability or confidence that the direction will be the specified one given that the "If" conditions are met. In this rule, the conditions appeared eight times. There is a confidence of 0.9 that the direction will be Up when these conditions occur.

Table 3. *Trained Model Rule Examples, 1 Per Year for Up and Down*

Training Year	Rule for Down	Rule for Up
1998	If JchgC <= -1.678 And HKDir = Up Then NDir = Down (6, 0.88)	If JchgC > -1.678 And JchgC <= -0.963 And Adays <= 1 And ADir = Up Then NDir = Up (8, 0.9)
1999	If JchgC <=-.0935 And HKDir = Up And AchgC <= 1.272 And Adays <=2 And ADir = Up Then NDir = Down (4, 0.83)	If AchgC > 1.272 Then NDir = Up (11, 0.923)
2000	If HKDir = Down And NDays <= 2 And AchgC > -0.108 And Adir = Down Then NDir = Down (7, 0.89)	If JchgC > -1.013 And Jchg C <= -0.605 And HKDir = Up And NDays <=2 Then NDir = Up (6, 0.88)
2001	If HKchgC > -1.284 And JchgC <= -0.438 And HKDir = Down And NDays <= 1 And Adir = Up Then NDir = Down (8, 0.90)	If JchgC > -0.864 And NDays <= 1 And AchgC <= -0.795 Then NDir = Up (5, 0.857)

Some rules are more complex and some are simple. Because a decision tree is built on data patterns which have occurred in the training set, we can use them to gain insight into trading actions. The rules put words to complicated responses from traders who often are unable to explain the complex reasoning behind their positions.

The percentage of correct daily direction forecasts of the Nasdaq for each of the yearly models is shown in Table 4. This data displays the percentage correct of each year in the corresponding validation set. As can be seen from Table 4, the decision tree forecasts work significantly better on days when the Nasdaq was Up

Table 4. *Percent of Correct Direction Forecasts*

Training Year	Forecasting Year	Correct Down	Correct Up
1998	1999	37.93	75.20
1999	2000	46.15	68.57
2000	2001	43.81	69.52
2001	2002	42.74	66.34

than on Down days. Other than the training year 1998, the percent of correct Down day forecasts was in the forty percent range. The Up day forecasts ranged from 66 to 75 percent correct.

Witnessing such results, especially in the context of previous studies on the topic of market interrelationships, naturally raises the question: is the observed leading effect a pure one, or are we in fact seeing the effect of Nasdaq leading both these Pacific Rim markets and itself—a situation which would lead to a correlation between markets that have no causal impact in the direction under consideration. In order to better understand this dynamic, decision trees were built to forecast movement of each of our leading markets based on the movement seen in the Nasdaq. That is, inputs of NchgC, NDays, and NDir on day t were used to forecast market direction in the Pacific Rim on day t+1. The results are below in Table 5.

Table 5. *Using Nasdaq to Forecast Australia, Japan, and Hong Kong.*

Training Year	Forecasting Year	Japan Down	Japan Up	Aust. Down	Aust. Up	HK Down	HK Up
1998	1999	46.94	64.91	54.46	72.07	68.42	56.41
1999	2000	87.40	37.89	80.77	45.76	78.95	43.52
2000	2001	57.89	60.42	40.00	83.48	62.28	65.63
2001	2002	94.02	29.70	72.58	69.15	64.80	59.14

Important to notice is the lack of consistency in the results shown in this table. While some years and markets provide uniformly better than random forecasting independent of direction (for example, HK in 1998/99, 2000/01 and 2001/02), other years and markets (especially Japan) give unreliable results from the standpoint of consistency, even as they predict some particular aspect of market performance extraordinarily well.

5. Conclusions

The above results imply a generally positive outlook. Where previous experiments have found little correlation between the movements of Pacific Rim markets and the subsequent changes in American markets, the use of data mining techniques has shown that in some cases, even very crude information can be used to make reasonable predictions about future performance. Not only is this information available, but the rules constraining its application are reasonably consistent over time: after training, the model performed reasonably and at a sustained level over a

full year's period.

Performance was especially good on days when the Nasdaq was up, with the decision tree making a correct forecast of direction on approximately seven out of ten days. On days in which the Nasdaq was down, however, performance was worse than chance, giving a correct forecast on only about four of ten days. The very encouraging results found on Up days imply that the information needed to make a reasonable prediction is present; perhaps with a more comprehensive set of data, additional rules to tighten the forecasting of Down days can be developed.
Also interesting is the complexity of the rules. As might be expected from the low correlations between market movements, we often find that an Down forecast for the Nasdaq is predicted when input markets are up by certain amounts, rather than the intuitive opposite. In addition, some rules incorporate constraints on multiple markets, or multiple constraints on the same market. In other words, predictive ability is sometimes realized when forecasts are based on the movements of multiple leading markets, or on multiple aspects of the same market, rather than on a single input.

Separating market movement into direction and change, rather than considering only the single numerical value representing said market movement, simulates the separation of this knowledge into two pieces of information: one, whether the market was up or down; and two, how much the market moved. This produces more favorable results, and very many of the decision rules refer explicitly to the direction of some market's movement, ignoring the magnitude of the change. Many rules also refer to the number of days since the market was last open, confirming the hypothesis that this is a worthwhile piece of information when formulating trading strategies.

At present, this model is not meant to be used as a predictive tool for playing the market successfully. Most strongly, it is a signal that there is ample information to be found where none has been located before, in the movements of markets that were once seen as purely lagging their counterpart American markets. Results are not enough to conclude a causal effect between the movements of the chosen Pacific Rim markets and the Nasdaq, but at the very least, with predictions correct at twenty percentage points above chance on Up days, there exists an undeniable relationship—though quite possibly, only that of a common source driving movements of all of the above markets. The argument that this common source is the Nasdaq itself—an argument which would imply that there are in fact no true leading effects shown in this paper—is made less plausible by the results in Table 5; significantly more reliable predictive ability in the inverse direction is not shown. Nonetheless, there is a good deal that remains to be done. Results, while significantly above chance in some cases, might stand to benefit from further study. Some suggestions for direction of future study in this area are addressed next.

6. Future Research Directions

This was a very limited study. Only three Pacific Rim markets were considered.

Even if we wish to restrict ourselves to markets which have opened and closed during the time a given target market is closed, there are many markets, both developed and emerging, that could be used as inputs into the decision tree. In addition, lags could be incorporated, as well as volatilities of the markets and other related variables. Thus, to summarize, additional information could be derived from other markets, from other days within the same markets, or from more specific information about the particular movements within a day; analysis of the information could be refined by application of a variety of data mining methods, especially cluster analysis. Cluster analysis is a preprocessing technique applied prior to using a decision tree. Its purpose is to extract groups of days with similar behavior. Use of cluster analysis can subdivide the data set into more uniform segments, so that particular decision rules can be tailored to more accurately predict the interrelationships within a given subset.

References

Berry, Michael, and Gordon Linoff, 2000, "Mastering Data Mining", Wiley.

Brooks, Robin, Kristin Forbes, and Ashoka Mody, 2003, "How Strong are Global Linkages?", www.imf.org/external/np/res/seminars/ 2003/global/pdf/over.pdf www.imf.org/external/np/res/seminars/ 2003/global/pdf/brook.pdf

Capelle-Blancard, B. and Helene Raymond, 2002, "Do International Stock Markets Linkages Change Across Bulls and Bears?", http://www.crereg.univ-rennes1.fr/seminaire-pages%20interieures/documents%20presentes%20en%20seminaires/CapelleRaymond1.pdf

"Clementine User Guide, Version 5", 1998, Integral Solutions Limited.

Darbar, S. and P. Deb, 1994, "Co-movements in International Equity Markets", Working Paper, Indiana University-Purdue University, Indianapolis.

Dwyer, G. and R. Hafer, 1988, "Are National Stock Markets Linked?", Federal Reserve Bank of St. Louis Report, November/December.

Fischer, K. and A. Palasvirta, 1990, "High Road to a Global Marketplace: The International Transmission of Stock Market Fluctuations", *The Financial Review*, Vol. 25, No. 3, August.

Groth, R., 1998, *Data Mining,* Prentice Hall.

Hand, David, Heikki Mannila, and Padhraic Smyth, , 2001, *Principles of Data Mining*, MIT Press.

Hilliard, J. 1979, "The Relationship Between Equity Indices on World Exchanges", *The Journal of Finance"*, Vol XXXIV, No. 1, March.

Malliaris, M. and L. Salchenberger, 2002, "Using Neural Networks to Discover Patterns in International Equity Markets: A Case Study" in *Neural Networks in Business: Techniques and Applications*, Smith and Gupta, eds., Idea Group Publishing.

Ribeiro, Ruy and Pietro Veronest, 2002, "The Excess Comovement of International Stock Markets in Bad Times", November, home.uchicago.edu/~rmribeir/cross.pdf