Dr. rer. nat. Hans UHLIG, Jagersredder 27a, 22397 Hamburg,
Tel. 040 / 605 51 50

The paper given below was submitted as shown to the American finance magazine 'Technical Analysis of Stocks & Commodities', Seattle, Washington, U.S.A. And published there in revised form in the September issue of 2001.

Summary

This paper introduces the so called 'nearest neighbour' prediction method as a prognosis instrument for financial markets. As an example the Standard & Poor's 500 stock index was chosen. This index usually serves as a benchmark for the performance of portfolio managers. The method uses the weekly closes of the index from January 1980 to December 2000 which are roughly 1000 data as a data base to predict the movement of the index for one week ahead. The method was applied for half a year, which made 26 predictions. The aim was to predict up or down movements correctly.

The possible returns using this method were compared to returns using the simple 'buy and hold' strategy. It was found that the 'nearest neighbour' predictor performed much better than the simple strategy. While the index lost considerably in the course the observation period the predictor was able to make a small profit.

Performance is not the only advantage of the 'nearest neighbour' method, though an important one. The particular appeal of the method is its transparency. There is no black box decision. And it is inexpensive because only an ordinary spread sheet program is needed for calculations and sorting of data sets. The spread sheet program of the OpenOffice(TM) program suite provides all the necessary functions and more than that.

by Hans Uhlig, Ph. D.
for 'Technical Analysis of STOCKS & COMMODITIES" (Sept. 2001)

Nearest Neighbour prediction

If we looked at the stock market in the last decade it was quite boring. It was a one way street, leading upwards. In those days, one of the best investment strategies was: 'buy and hold'. This in historical perspective quite abnormal market behaviour has lead to wide spread overestimation of the 'buy and hold' strategy and at the same time to an underestimation of market timers. Meanwhile the picture has changed somewhat. Indices are not climbing to new highs anymore, but are rather testing their lows. In these times 'buy and hold' is not the investment strategy of choice as the unbiased reader will certainly admit. A look at other approaches to stock market investment thus seems advisable.

Here I will show you how chaos theory can help us to make better investment decisions. I will introduce a method borrowed from chaos research, which is called 'k nearest neighbour prediction', named NNP for short in the following. Using terms of 'technical analysis' the NNP method is best described as a kind of automated chart interpretation. Based on weekly closes of the S&P 500 the NNP will be used to make one week forecasts of the stock market direction. Investing accordingly, would clearly have outperformed 'buy and hold' over the last five years, even at times when 'buy and hold' was a good choice. The beauty of the method is manifold: it requires only an ordinary spreadsheet program. The operating principle is easy to grasp. Decisions are always transparent, as opposed to neural nets. There is no lengthy training phase. The method is quite robust and even self improving over time, the more data become available.

Local versus global predictors

The NNP is a local predictor, in contrast to global predictors, most of you will be familiar with. A method which averages over all available data to make estimates is a global method. Auto Regressive Integrated Moving Averages (ARIMA) are examples of global predictors. 'Technical analysts' , however, look if the actual chart shows a certain significant pattern, which resembles similar situations in the past. Only similar situations are considered for interpreting the chart. Predictions based on chart patterns are examples of local predictors.

Ordinary chart interpretation and NNP - similarities and differences

Chart reading has two critical aspects: first, it is a visual method and similarity is in the eye of the beholder, second, at times a chart does not show a significant pattern. The NNP approach is a systematical one. First we have to define a pattern, then we define how to quantify similarity between any two patterns. Having done this we can sort patterns according to their similarity to a reference pattern, which in most cases will represent the most recent market situation. Finally we look at the most similar patterns and how the market developed in these cases. The average of the moves

that followed the nearest neighbours' patterns is an estimate of the move to be expected after the reference pattern occurred. Patterns here there and everywhere, but what is a pattern? How many and which data could build a suitable pattern? This is where non linear statistics and chaos theory come in.

What chaos theory can teach us

Chaos theory says that the complete dynamics of a chaotic multi component system can be reconstructed from a time series of a single component. In the case of a market the price would be the component to be used to reconstruct the complete dynamics. This means, we need only price data to build our patterns. One question remains: is a market a chaotic multi component system? The answer is: probably. Since markets can look like random, hence the 'random walk' hypothesis, they are either random or non linear (chaotic). Of course nobody would doubt that randomness is involved in market movements occasionally, but the inherent dynamics are certainly non random as can be shown beyond doubt with well established tests and some new methods developed by chaos researchers.

Preserving information contained in the data

Two tests of non linear relationships between data are particularly useful: the chi² - test of independence and the test of conditional entropy, both of which are well known to statisticians, but you will probably not find them in your spreadsheet program. With a chi² - test of independence we can find out that the stock market has a memory and how significant this memory is. And with conditional entropy measurements we can determine how much information about the future is contained in the present and past data and how this information declines with time. Knowing about this can help us decide, which and how many data have to be included in a single pattern in order to preserve enough information. If you do not have access to programs providing this type of non linear statistics, which will be the rule, you will have to find out suitable data by trial and error. If you are lucky and have the tests available, you should bear in mind that they use global methods. They are not tailored to local predictors, but can only give rough and ready estimates.

Defining a pattern

Earlier testing had shown that it was favourable to work with the natural logarithm of the index and that four data would suffice to represent a stock market situation at any time. For the S&P we choose the weekly differences of the (natural) log of closes of one, two, four and five weeks earlier than the week to be predicted as our pattern.

Measuring similarity

Similarity is measured by calculating the absolute difference (Manhattan metric, city block norm) of all other patterns relative to the reference pattern (see the formula in H16 of the screen shot). The abs(...) means that each individual difference receives a positive sign, this prevents upside and downside differences from summing up to zero. The $ signs in the formula denote an absolute relation. If we forgot them, only one pattern would be compared to the reference pattern and all others would be compared to one another. What about other distance measures, you might ask. They do exist, and will be discussed below.

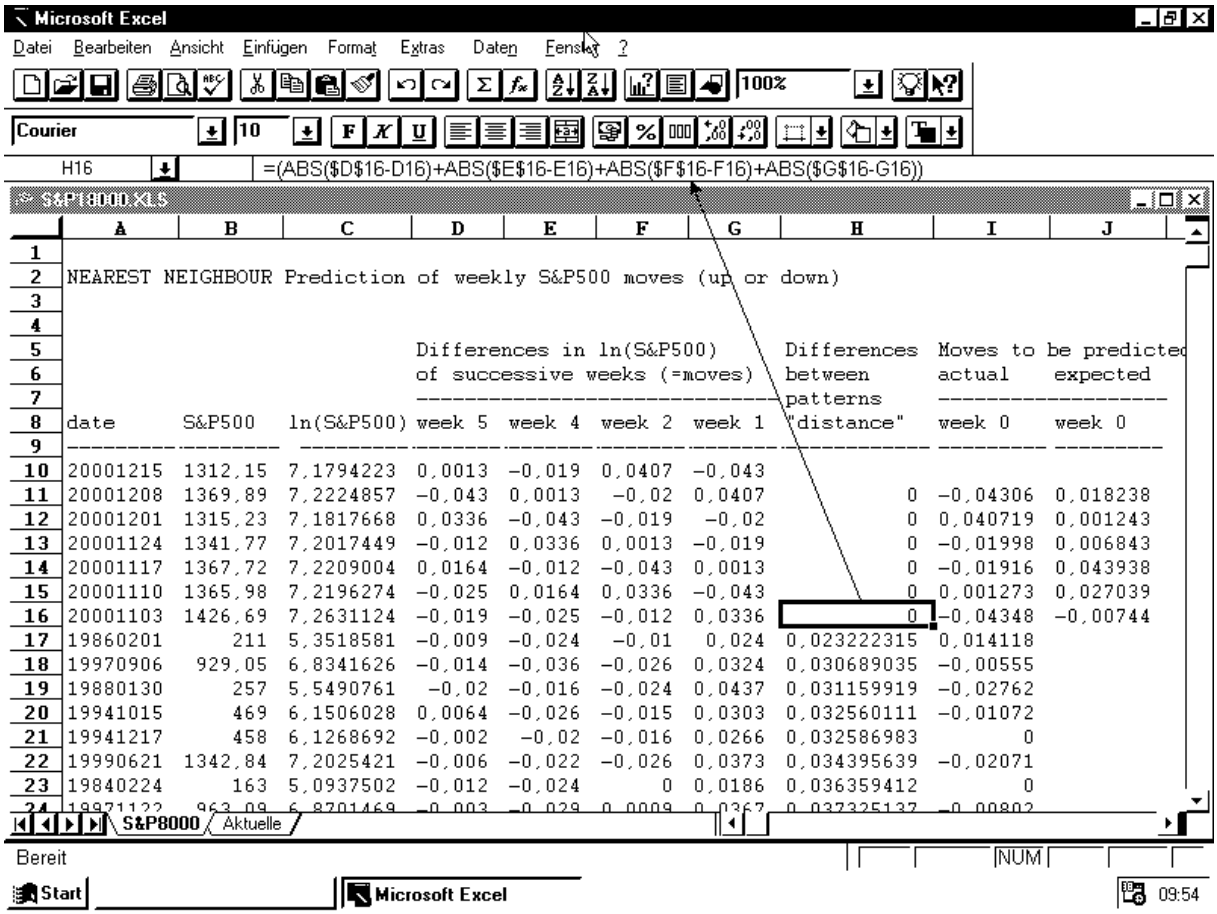


Figure 1 - Spreadsheet with NNP method, the description is given in the text.

Sorting of patterns over distance

Once the distances of all data sets (patterns) in the database relative to the reference data set (pattern) are calculated, the data sets are sorted according to their differences to the reference data set. The data sets being closest to the reference set are the nearest neighbours. We know how the nearest neighbours developed, that is what their next move was. All we have to do now is to average over the next moves of the nearest neighbours and we have an estimate of what the movement of our reference data set will be. We have found that four nearest neighbours are a good choice to start with. Thus the k in 'k nearest neighbours prediction' equals four.

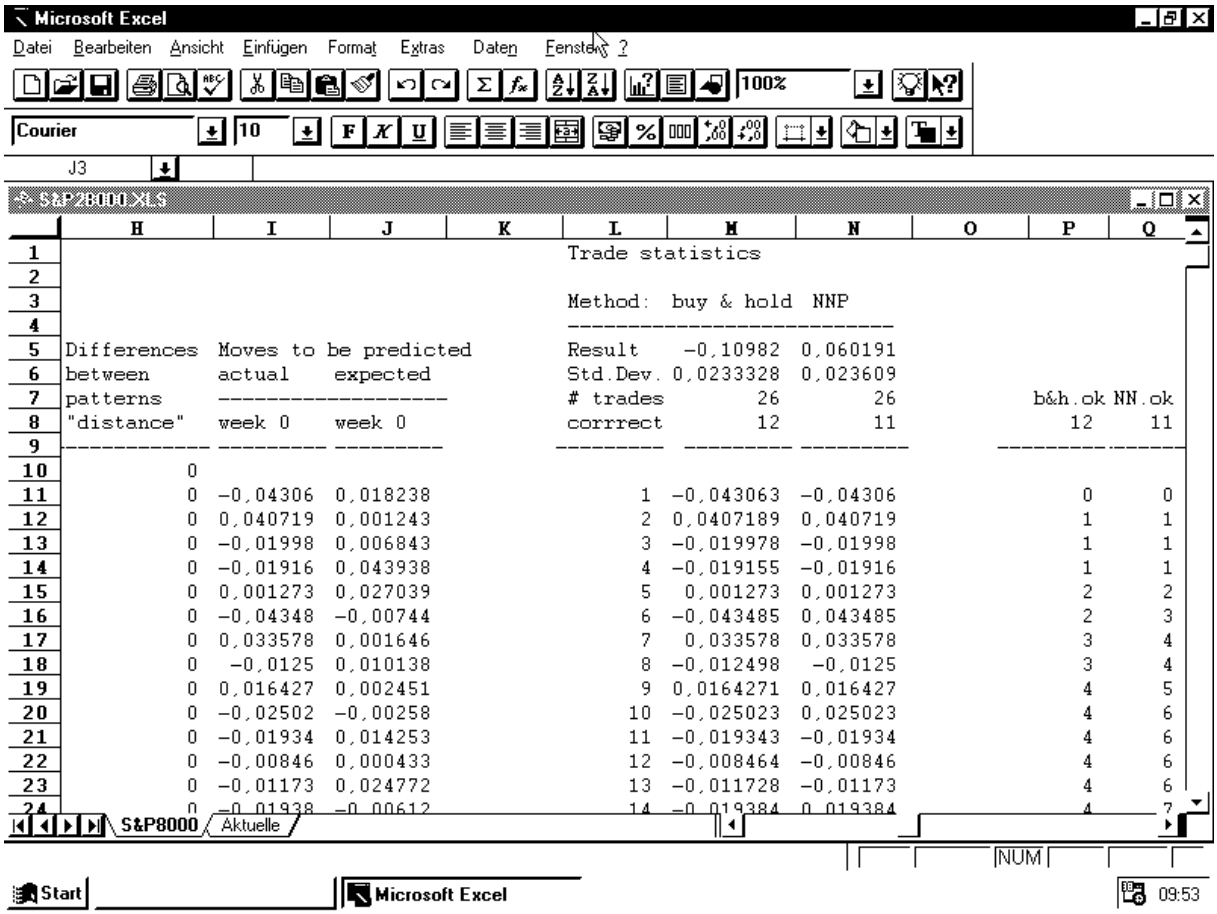
NNP of the S&P500

The first screen shot shows an instructive example prediction for the S&P500. As the data base I used weekly closes of the time range January 1980 to December 2000, which comprise roughly 1100 data. The patterns used are given in columns D through G. Column I shows the move which followed the respective pattern. It is not part of the pattern and not used to calculate distance. Column I is needed to make an estimate of the next move, for the reference pattern. The marked cell H16 holds the formula for the distance calculation. This formula was copied to column H of all the data sets prior to Nov 3, 2000 and then the data sets (patterns) were sorted according to their distance in ascending order, the most similar pattern being in row 17 and the next three nearest neighbours in rows 18 through 20. Column A shows when the patterns did occur: 1986, 1997, 1988 and 1994. The very nearest

neighbour was from Feb 1 in 1986, but at that time the next move was up, as we see in cell I17. The next three neighbours, however, all went down. So the average of the cells (I17:I20) is a downward move of -0.0074 as computed in cell J16 this shows the advantage of using several nearest neighbours. See also discussion of robustness below.

Results

The method described above is the most simple form of the NNP. But if done properly it works without further refinement and can outperform 'buy and hold'. I have carried out the NNP for the last 26 weeks with a spreadsheet program and found that 'buy & hold' was the right decision 12 times but overall it lead to a small loss, while NNP was right only eleven times but would have made a small gain. The standard deviations of results were almost the same. Such a small number of tests can hardly be called representative, but on the other hand it would be very laborious to make a huge number of tests with a spreadsheet program. I have looked at the performance of NNP over longer periods up to 12 years using an NNP program I have coded. Some printouts of reports comprising the last 5 years given by that program for different parameter settings (number of nearest neighbours, pattern structure, distance weighting, input weighting) are shown below. The problem with longtime studies is that the data base becomes smaller and smaller the deeper you look into the past, because only data older than the reference pattern can be used for nearest neighbour predictions.

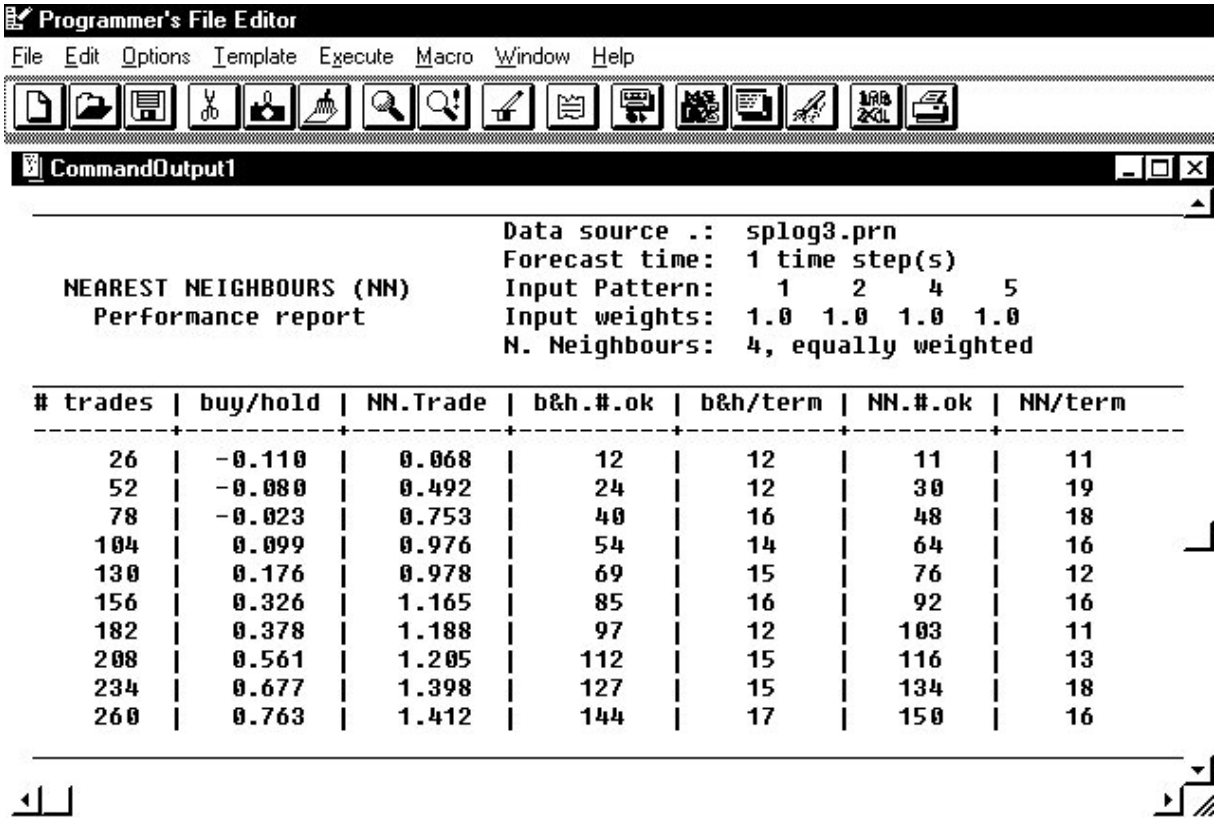


	H	I	J	K	L	M	N	O	P	Q
1					Trade statistics					
2										
3					Method:	buy & hold	NNP			
4										
5	Differences	Moves to be predicted			Result	-0,10982	0,060191			
6	between	actual	expected		Std.Dev.	0,0233328	0,023609			
7	patterns				# trades	26	26		b&h.ok	NN.ok
8	"distance"	week 0	week 0		corrrect	12	11		12	11
9										
10	0									
11	0	-0,04306	0,018238		1	-0,043063	-0,04306		0	0
12	0	0,040719	0,001243		2	0,0407189	0,040719		1	1
13	0	-0,01998	0,006843		3	-0,019978	-0,01998		1	1
14	0	-0,01916	0,043938		4	-0,019155	-0,01916		1	1
15	0	0,001273	0,027039		5	0,001273	0,001273		2	2
16	0	-0,04348	-0,00744		6	-0,043485	0,043485		2	3
17	0	0,033578	0,001646		7	0,033578	0,033578		3	4
18	0	-0,0125	0,010138		8	-0,012498	-0,0125		3	4
19	0	0,016427	0,002451		9	0,0164271	0,016427		4	5
20	0	-0,02502	-0,00258		10	-0,025023	0,025023		4	6
21	0	-0,01934	0,014253		11	-0,019343	-0,01934		4	6
22	0	-0,00846	0,000433		12	-0,008464	-0,00846		4	6
23	0	-0,01173	0,024772		13	-0,011728	-0,01173		4	6
24	0	-0,01938	-0,00612		14	-0,019384	0,019384		4	7

Figure 2 - Spreadsheet with NNP trade statistics table

The 'trade' statistics is given in columns L ..N, rows 1...8. Here you see a concise performance report of NNP as compared to 'buy and hold', including standard deviations of returns, which are almost identical, i..e. differ in the 4th decimal digit.

You may wonder if the method can be applied to other stock market indices, to individual stocks and to other financial markets in general. Well, to be honest, the S&P500 belongs to the easier tasks, because of its breadth. A good pattern definition may look quite different for other markets and the same holds true for other parameters. The NNP can be successfully applied to estimate moves in other international stock indices (Nikkei 225, FTSE 100) and even the crude oil price direction or the Yen/US-$ exchange rate changes over the past 5 years, to name examples from the commodities and currencies markets. But the NNP it is by no means limited to these and it can also beat other strategies than 'buy and hold'.



# trades	buy/hold	NN.Trade	b&h.#.ok	b&h/term	NN.#.ok	NN/term
26	-0.110	0.068	12	12	11	11
52	-0.080	0.492	24	12	30	19
78	-0.023	0.753	40	16	48	18
104	0.099	0.976	54	14	64	16
130	0.176	0.978	69	15	76	12
156	0.326	1.165	85	16	92	16
182	0.378	1.188	97	12	103	11
208	0.561	1.205	112	15	116	13
234	0.677	1.398	127	15	134	18
260	0.763	1.412	144	17	150	16

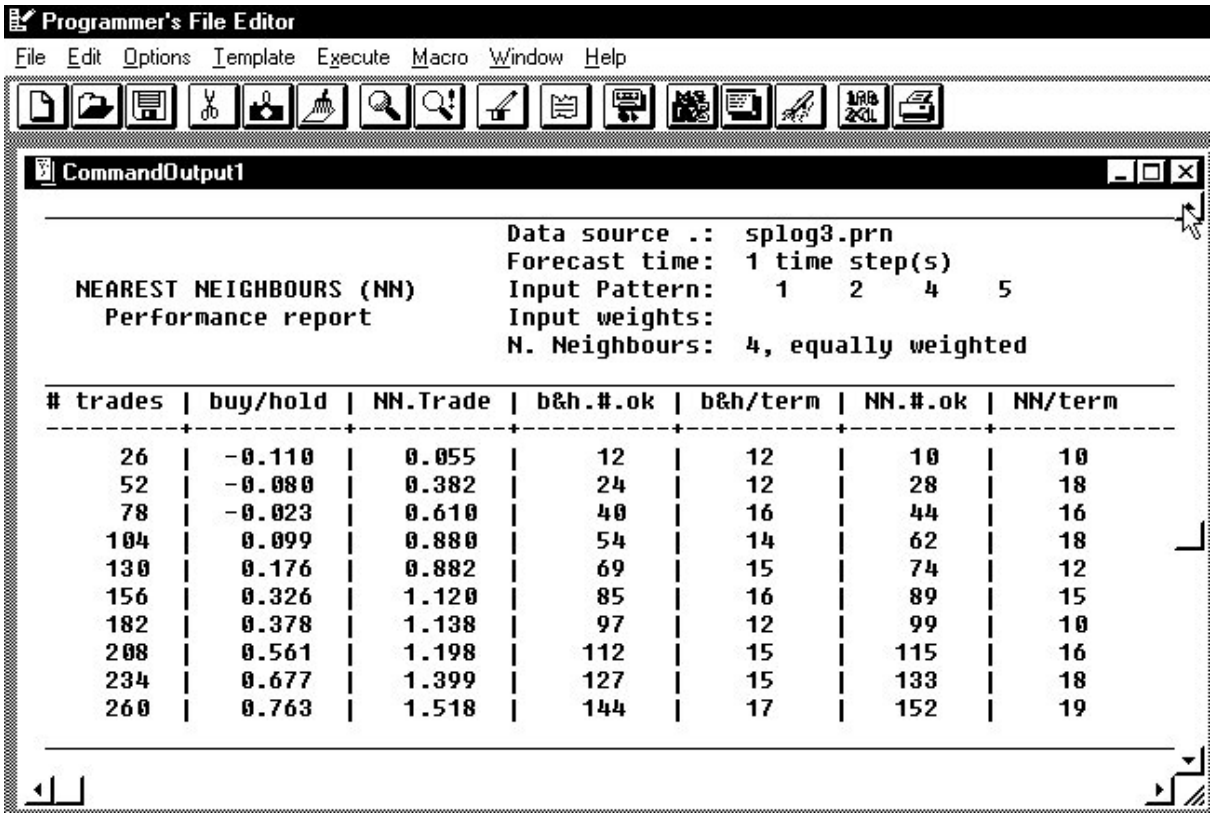
Figure 3 - Output of NNP program

'Performance report' for the NNP with parameter settings identical to those used in the spreadsheet, but carried on for the past 5 years. The columns from left to right, far left: the number of 'trades', that is weekly decisions; next the results for 'buy and hold', next results for NNP, next how often 'buy and hold' was the right decision relative to the number of trades given in the column far left; next how often 'buy and hold' was the right decision for that half year period. The next two columns give the same data as just described, but for the NNP method. The rows from top to bottom give the cumulated results for the number of trades except for the columns headed 'b&h./term' and 'NN/term'

If you compare the statistics table of screen 2 and the first row of this table, there is a slight discrepancy (less than 1%) between the two outputs. This is due to the fact that in the spreadsheet the logs were calculated and used with the full number of eighteen decimal digits, while in the data file used for the NN program the ln(index) values had only six decimal digits. The inevitable round off error is usually quite small, but can at times even lead to different neighbourships and thus possibly big differences, though these will average out in the long run.

Robustness
The method seems to be quite robust, meaning noise tolerant and insensitive to small changes. Robustness has several reasons. Four data build each pattern, so that abnormal individual values cannot distort a pattern too much. Using the absolute difference as the distance measure also helps to keep the method robust (see below). The final decision is made by averaging over four nearest neighbours, which again reduces the weight of potential outliers.

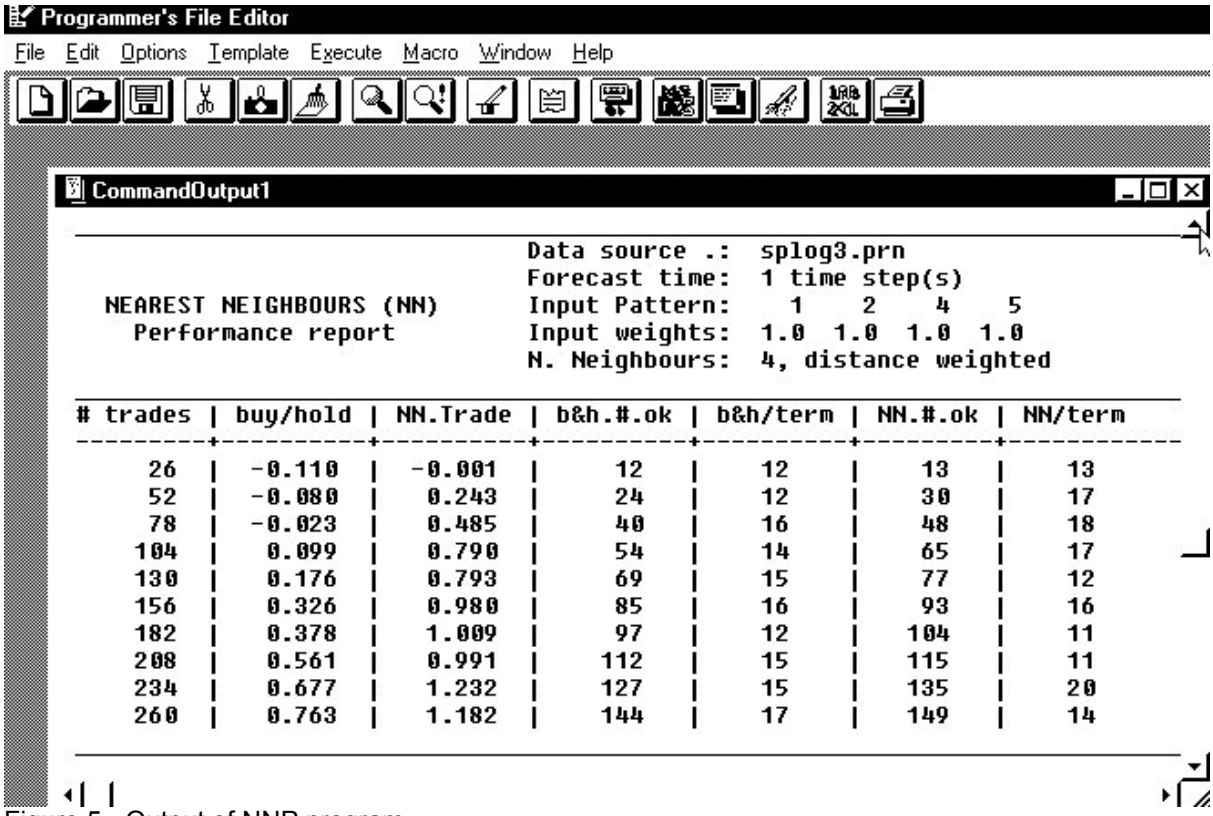
Refinements
Several refinements are known and they really can add to the performance as will be shown. The first refinement applies to the weighting of the nearest neighbours. So far we have taken a simple average over the four nearest neighbours. This means the closest neighbour has the same weight as the fourth neighbour. The method can be improved by weighting each neighbour proportionally to its inversed distance. That is, the smaller the distance of a neighbour, the more weight it gets. Weightings have to be normalized (each weight devided by the total weight) of course.



# trades	buy/hold	NN.Trade	b&h.#.ok	b&h/term	NN.#.ok	NN/term
26	-0.110	0.055	12	12	10	10
52	-0.080	0.382	24	12	28	18
78	-0.023	0.610	40	16	44	16
104	0.099	0.880	54	14	62	18
130	0.176	0.882	69	15	74	12
156	0.326	1.120	85	16	89	15
182	0.378	1.138	97	12	99	10
208	0.561	1.198	112	15	115	16
234	0.677	1.399	127	15	133	18
260	0.763	1.518	144	17	152	19

Figure 4 - Output of NNP program
This output differs from the previous one by the fact that inputs were not weighted uniformly, but individually.

A further refinement would be, to apply different weights to the individual data building each pattern. To many of us it seems logical that the most recent move of the market may be more important for the next one than the change having occurred say three weeks ago. But you will be surprised: The recent move is not as important as you think and earlier moves are more important than you - and the 'random walk' proponents - would believe. Nevertheless individual weights do improve the performance of the method.



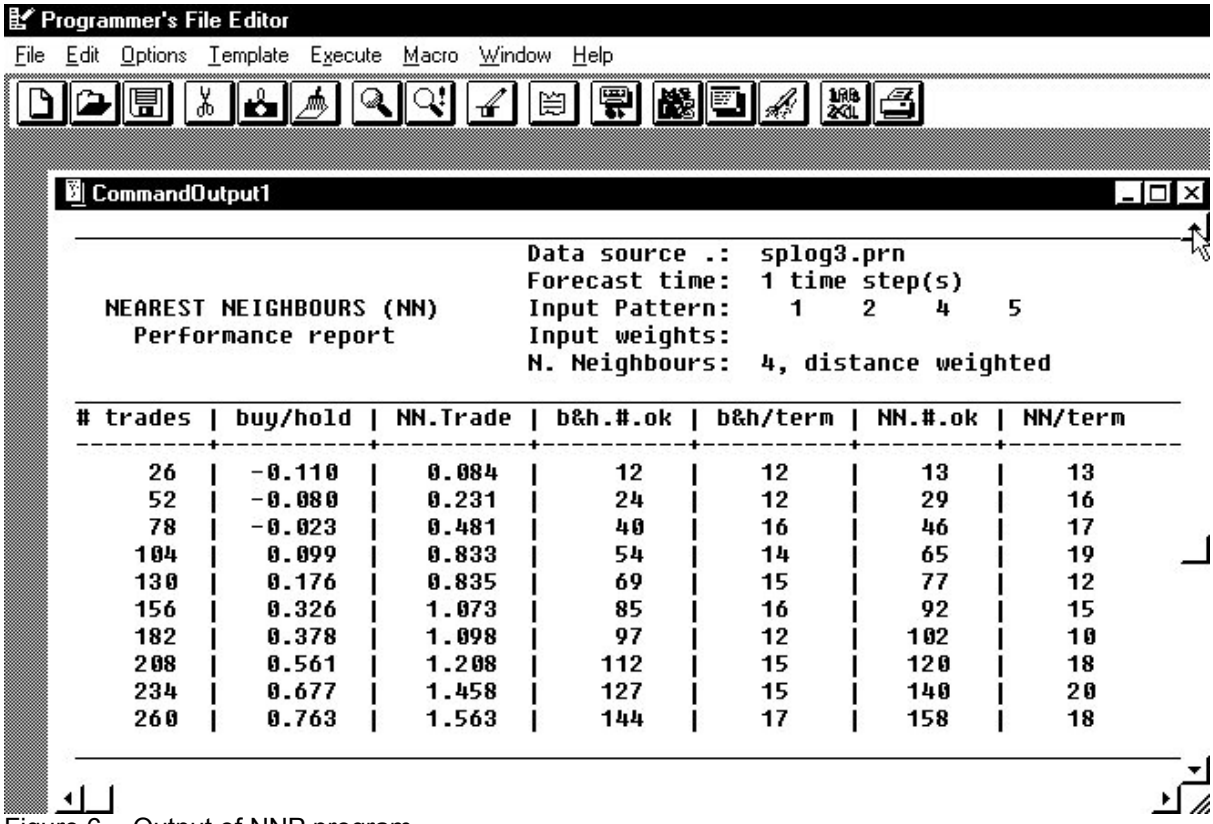
# trades	buy/hold	NN.Trade	b&h.#.ok	b&h/term	NN.#.ok	NN/term
26	-0.110	-0.001	12	12	13	13
52	-0.080	0.243	24	12	30	17
78	-0.023	0.485	40	16	48	18
104	0.099	0.790	54	14	65	17
130	0.176	0.793	69	15	77	12
156	0.326	0.980	85	16	93	16
182	0.378	1.009	97	12	104	11
208	0.561	0.991	112	15	115	11
234	0.677	1.232	127	15	135	20
260	0.763	1.182	144	17	149	14

Figure 5 - Output of NNP program
This output shows distance weighted nearest neighbours but is identical to screen 3, otherwise.

There is much room left for further improvements of the NNP as described here. So far we have used only a small fraction of the available data. Daily quotes could be used and longer time series could further add to the performance. You may perhaps think of another distance measure, like Euklidian metric (root of the sum of squared differences). I have tried it, but it was not as satisfactory as the Manhattan metric. First, the ratio of right:wrong decisions was lower. And because of the second order calculations involved in the Euklidian metric the results are less stable, leading to greater performance deviations in consecutive half year periods.

A word of caution

The NNP is no crystal ball which reveals the future. Even if well constructed it has been and will be wrong in many cases. In the long run the right:wrong ratio of decisions was between 55:45 and 60:40 and the NNP seemed to make several 'big points'. But of course this is no guarantee for the future. Just using NNP will certainly not make you a successful trader. I can only repeat and emphasize the messages given by many experienced stock market connaisseurs in the interviews published in this journal, that you have to do your homework. You must have a trading system and you must have a money management system. And last not least you must have a trader's self discipline, which as to 'behavioral finance' is quite a rare gift.



# trades	buy/hold	NN.Trade	b&h.#.ok	b&h/term	NN.#.ok	NN/term
26	-0.110	0.084	12	12	13	13
52	-0.080	0.231	24	12	29	16
78	-0.023	0.481	40	16	46	17
104	0.099	0.833	54	14	65	19
130	0.176	0.835	69	15	77	12
156	0.326	1.073	85	16	92	15
182	0.378	1.098	97	12	102	10
208	0.561	1.208	112	15	120	18
234	0.677	1.458	127	15	140	20
260	0.763	1.563	144	17	158	18

Figure 6 - Output of NNP program
An example of distance weighted neighbours + weighted inputs.

Here I frequently made reference to chaos theory, which is the popular term for the theory of complex dynamic systems. The basis of which is topology, a branch of mathematics dealing with those properties of geometric figures which remain constant under transformation. To the public it is known as 'rubber mathematics'. These provide the tools and the scientific background for using the NNP. By establishing consistent rules for interpreting market patterns and thus taking away the subjectivity so far involved in chart interpretation, the NNP is not merely a 'technical' analysis, but could be considered a scientific method. However, do not demand too much. Markets are social systems and not mathematical equations or physical systems, like those dealt with in chaos theory. They do not follow eternal laws, but rather obey to rules and at times, it seems, not even this. One reason is that noise from outside the market will transiently disturb the dynamics. Another reason is that markets adapt to a changing environment, like all living systems, developing towards more complicated rules. Therefore attempts to make mathematical models of markets can only have limited success.

Further reading

Here you only got a glimpse of what chaos theory might have to offer for market analysts and interested investors. For more information you could look at the author's INTERNET site at http://www.hans-uhlig.de or you could email to Hans.Uhlig@hamburg.de